Compression of Multi-View-plus-Depth (MVD) data: from perceived quality analysis to MVD coding tools designing

Emilie Bosc

THESE INSA Rennes
sous le sceau de l'Université européenne de Bretagne
pour obtenir le titre de
DOCTEUR DE L'INSA DE RENNES
Spécialité : Traitement du signal et des images

présentée par
Emilie Bosc

ECOLE DOCTORALE : MATISSE
LABORATOIRE : IETR

Compression des données Multi-View-plus-Depth (MVD): de l'analyse de la qualité perçue à l'élaboration d'outils pour le codage des données MVD

Thèse soutenue le 22.10.2012
devant le jury composé de :

Azeddine Beghdadi
Professeur des universités, Université de Paris 13, France/ Président
Frédéric Dufaux
Directeur de recherche au CNRS, LTCI, Paris, France / rapporteur
Lina Karam
Professeur, Arizona State University, USA / rapporteur
Mårten Sjöström
Enseignant-chercheur, Mid Sweden University, Suède / examinateur
Quan Huynh-Thu
Ingénieur de recherche, Technicolor R&D, France / examinateur
Patrick Le Callet
Professeur des universités, Polytech'Nantes, France/ examinateur
Muriel Pressigout
Maître de conférences, INSA de Rennes, France / Co-encadrante de thèse
Luce Morin
Professeur des universités, INSA de Rennes, France / Directrice de thèse

Contents

Remerciements

Durant ces trois années de thèse, j'ai été entourée, choyée et soutenue par de nombreuses personnes. Je leur dédie humblement cette page.

Je remercie les membres du jury pour leur disponibilité et pour leurs conseils.

Je ne remercierai jamais assez les deux équipes qui m'ont accueillie durant cette thèse. Il s'agit du Groupe Image de l'Institut d'Electronique et de Télécommunications de Rennes (IETR) ainsi que du Groupe Image and Video Communications (IVC) de l'Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN). J'ai bénéficié dans ces deux équipes d'un accueil chaleureux et d'une sérieuse expertise et cette thèse en est le résultat.

Je remercie ma directrice de thèse, Luce Morin, pour ses conseils, son soutien, le temps qu'elle a passé à répondre à mes sollicitations, ses nombreuses qualités, son humanité et sa confiance. Merci également à Muriel Pressigout, qui a co-encadré cette thèse, pour tous ses conseils. Je remercie aussi Patrick Le Callet dont le rôle dans cette thèse a été très important puisqu'il m'a tout d'abord accueillie à l'IVC et m'a offert de nombreuses opportunités pour enrichir mes projets de recherche durant ces trois ans.

Je remercie mon premier mentor, Jonathan Piat, doctorant à l'IETR lors du début de ma thèse, qui m'a prodigué de nombreux conseils concernant la gestion de la vie trépidante de "thésard nomade". Je pense aussi à Cong Bai, Youssef Alj, Josselin Gautier, Fabien Racapé, Aurore Arlicot, Junle Wang, Jingpeng Li, Dalila Goudia, Sofiance Medjkoune, Cédric Ramassamy, Jing Li et Lu Zhang qui ont également été très présents pendant ces trois ans, et qui ont su me rassurer durant les épisodes difficiles et durant les moments de joie.

J'adresse à Romulad Pépion des remerciements particuliers pour sa disponibilité et son professionnalisme.

Je n'oublie pas non plus les petites mains qui m'ont guidées dans les labyrinthes administratifs et m'ont facilité bien des démarches : je pense à Mesdames Jocelyne Trémier, Corinne Calo et Frédérique Dessoliaire.

Enfin, je souhaite remercier ma famille pour son soutien : mes parents ainsi que mes frères qui m'ont toujours soutenue malgré la distance. Enfin, je remercie encore plus particulièrement mon conjoint qui pendant trois ans m'a soutenue malgré la distance parfois, et qui dans la mesure du possible, pendant trois ans, m'a conduite à la gare chaque matin à l'aube, et était présent sur le quai chaque soir, peu importe les retards de trains, et n'a cessé de me choyer. Je dédie cette thèse à mon conjoint sans qui l'éclat de ce travail aurait été totalement différent.

Résumé en français

1.1 Introduction générale

Les vidéos 3D sont considérées comme l'évolution de la télévision conventionnelle actuelle. Le changement radical attendu est comparé à celui qu'occasionna l'introduction de la couleur à la télévision. Dans le cas des vidéos 3D, l'innovation majeure vient de l'apport de l'impression de profondeur générée soit par l'exploitation du phénomène de stéréopsie (perception de la profondeur de champ relative de deux stimuli présentés dans le champ visuel) soit par le phénomène de parallaxe grâce à la navigation libre dans la scène.

Ainsi, les représentations telles que les vidéos multi-vues ((*Multi View Video* en anglais, il s'agit de l'ensemble de vidéos de couleur uniquement, à différents points de vue de la même scène, noté MVV) permettent la création de vidéos 3D. Il s'agit de plusieurs séquences vidéo conventionnelles prises avec plusieurs caméras synchronisées et à des positions différentes dans la scène. Lorsque l'on associe ces vidéos à des vidéos dites de profondeur on parle de données *Multiview Video-plus Depth*, MVD. La Figure 1.1 illustre ce type de données constitué de séquences en couleur et de séquences de profondeur.

La connaissance de la géométrie de la scène (issue des vidéos de profondeur) facilite la génération d'images virtuelles selon des points de vue différents de ceux réellement acquis par les caméras. A partir d'au moins une vue de couleur et de sa profondeur associée, on peut générer une vue virtuelle, c'est-à-dire non acquise par les caméras réelles, grâce aux algorithmes de synthèse. Ces algorithmes sont désignés par l'abréviation DIBR[Feh04], pour Depth-Image-Based-Rendering. Dans ce document, on désigne également par "vues synthétisées" les vues virtuelles générées par ces algorithmes.
Ces représentations de données permettent des applications telles que la télévision tridimensionnelle (3DTV) et le libre choix du point de vue (*Free viewpoint video*, FVV). La TV 3D donne une impression de profondeur ou de relief à la scène alors que la FVV offre à l'utilisateur la possibilité de choisir interactivement un point de vue arbitraire.

Les données MVD [SMS+07] sont de taille considérable et leur compression constitue un enjeu majeur pour les chercheurs s'intéressant à cette question. Cette thèse adresse cette problématique de compression des vidéos multi-vues avec pour pilier un souci constant du respect de la perception humaine du media. Les études et les choix portés durant cette thèse se veulent orientés par la recherche de la meilleure qualité perçue possible des vues synthétisées.

Puisque les MVD contiennent des informations de la même scène, à des points de vue

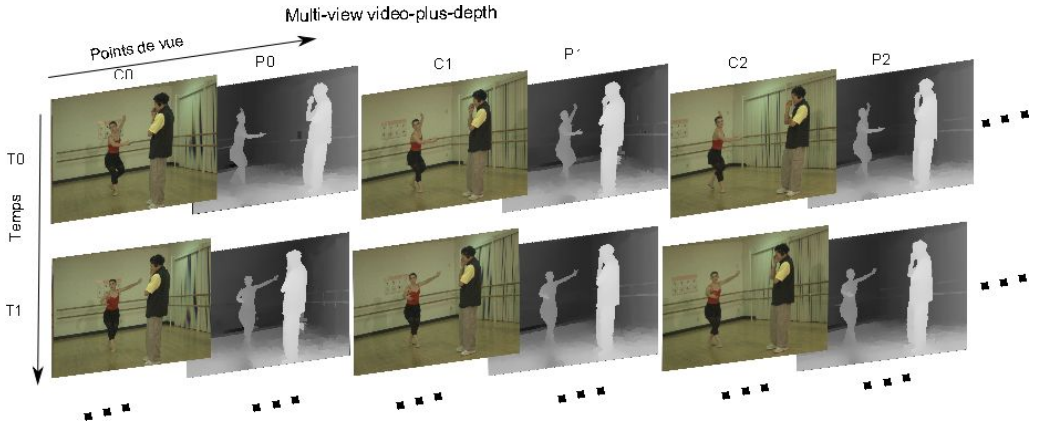


Figure 1.1 – *Données MVD. Les données de type MVD comportent des séquences de couleur (C0, C1 et C2) et de profondeur (P0, P1 et P2).*

légèrement différents, l'idée émergente [MSMW07b] consiste à exploiter autant que possible les redondances de ces données, d'une vue à l'autre, voire entre la couleur et la profondeur associée d'une même vue. A ce jour, il n'existe pas de méthode de compression de référence pour les MVD. La plupart des méthodes proposées se veulent des extensions du standard très connu H.264 AVC [MMSW06]. Les nombreuses propositions reposent sur l'utilisation de ce standard soit pour la compression des séquences de couleur et de profondeur séparément, soit uniquement des séquences de couleur. En effet, concernant la compression des séquences de profondeur, des études récentes [MMS+09] ont révélé que l'application de méthodes originellement conçues pour la compression de la couleur, infligent des dégradations aux cartes de profondeur. Ces dégradations entraînent des distorsions perceptibles et gênantes à l'issue d'une synthèse de vue à partir de ces séquences de profondeur décodées.

Tout l'enjeu des travaux que nous entreprenons réside dans l'investigation de nouvelles techniques de compression des données MVD limitant autant que possible les dégradations perceptibles sur les vues synthétisées à partir de ces données décodées. La difficulté vient du fait que les sources de dégradations des vues synthétisées sont d'une part multiples et d'autre part difficilement mesurables par les techniques actuelles d'évaluation de qualité d'images. Pour cette raison, les travaux de cette thèse s'articulent autour de deux axes principaux : l'évaluation de la qualité des vues synthétisées ainsi que les artefacts spécifiques et l'étude de schémas de compression des données MVD aidée de critères perceptuels.

1.2 Contributions

1.2 Evaluation de la qualité des vues synthétisées par DIBR

La synthèse de nouveaux points de vue est incontournable dans les deux applications principales de la vidéo 3D, que sont la TV 3D et la FVV. Or la qualité de cette synthèse est un facteur essentiel du succès de la vidéo 3D. Ainsi l'évaluation des vues synthétisées doit reposer sur des outils fiables et adaptés. Des expériences ont été réalisées dans le but de déterminer l'adéquation des méthodes d'évaluation de qualité reconnues pour les media conventionnels (subjectives et objectives) dans le cas des vues synthétisées par DIBR. Ces expériences ont été menées dans des conditions de visualisation monoscopique

et stéréoscopique.

Dans le cas des expériences en condition de vue monoscopique, les images fixes ainsi que les séquences vidéo ont été l'objet des tests. Dans le cas des expériences en condition de vue stéréoscopique, seules les images stéréoscopiques fixes ont été évaluées.

La plupart des méthodes proposées pour l'évaluation des media 3D reposent sur l'extension de métriques traditionnellement appliquées aux media 2D. De précédentes études ([YHFK08, TGSM08, HWD+09]) ont estimé la fiabilité de ses métriques objectives de qualité d'images. Dans [YXPW10], You *et al.*ont étudié l'évaluation de la qualité des paires stéréoscopiques en condition de vue stéréoscopique en utilisant des métriques 2D, mais les paires testées n'incluaient pas de distorsions liées aux DIBR. Dans de nombreuses études de ce type, les protocoles expérimentaux impliquent souvent à la fois des artefacts de compression de la carte de profodeur et des artefacts de compression des images de couleur associées, différents types d'écrans 3D, et différentes représentations de vidéos 3D (2D+Z, stéréoscopique, MVD, etc...). Dans ces cas d'étude, les scores de qualité subjective obtenus sont alors comparés aux scores de qualité objective, dans le but de trouver une corrélation et de valider la pertinence de l'utilisation des métriques 2D testées. Cependant, les artefacts liés au processus de synthèse et ceux liés à la quantification des données de couleur et de profondeur sont mesurés en même temps et sans distinction.

Les expériences réalisées dans cette étude concernent les vues synthétisées par différents algorithmes DIBR, à partir de données n'ayant pas subi de compression. Ces vues sont observées en monoscopique (sur un écran conventionnel), et en stéréoscopique (sur un écran stéréoscopique).

Protocoles expérimentaux

Trois séquences MVD ont été utilisées. Il s'agit de Book Arrival (1024x768, 16 caméras espacées de 6.5 cm), Lovebird1 (1024x768, 12 caméras espacées de 3.5 cm) et Newspaper (1024x768, 9 caméras espacées de 5 cm). Sept algorithmes DIBR ont chacun permis la génération de quatre nouveaux points de vue, pour chaque séquence. On leur assigne une étiquette allant de $A1$ à $A7$:

- A1 : repose sur l'algorithme présenté dans Fehn [Feh04], la carte de profondeur est pré-traitée par un filtre passe-bas. Les bords de l'image sont coupés et une méthode d'interpolation permet d'atteindre la taille originale de l'image.
- A2 : repose sur l'algorithme présenté dans Fehn [Feh04]. Les bords de l'image sont extrapolés par *inpainting* avec la méthode de [Tel04].
- A3 : Tanimoto *et al.* [MFY+09], est la méthode utilisée en tant que software de référence dans les expériences du groupe 3DV de MPEG.
- A4 : Müller *et al.*[MSD+08], propose une méthode de remplissage aidée de l'information de profondeur.
- A5 : Ndjiki-Nya et al [NNKD+10], propose une méthode de remplissage basée sur des patchs de synthèse.

- A6 : Köppel *et al.*[KNND+10], utilise l'information de profondeur dans le domaine temporel pour améliorer la synthèse des zones découvertes.
- A7 : correspond aux séquences pour lesquelles les zones découvertes ne sont pas remplies (donc avec des trous).

Les tests ont été conduits suivant les recommandations de l'ITU. Pour les évaluations de qualité subjective en condition monoscopique, les stimuli ont été présentés sur l'écran TVLogic LVM401W, selon ITU-T BT.500 [BT.93]. Pour les évaluations de qualité subjective en condition stéréoscopique, les stimuli ont été présentés sur l'écran Acer GD245HQ screen, avec NVIDIA 3D Vision Controller. Dans la suite, les méthodes d'évaluation subjective sont présentées puis les métriques objectives sont définies. Les mesures objectives de qualité ont été obtenues grâce à l'outil MetriX MuX Visual Quality Assessment Package [Mux].

Méthodologies d'évaluation subjective de qualité d'images

En l'absence de méthode d'évaluation adaptée aux conditions particulières de la 3D, l'évaluation des vues synthétisées repose sur des méthodes reconnues d'évaluation de media 2D. Le tableau 1.1 répertorie les méthodologies couramment utilisées pour l'évaluation des media 2D. Nos expériences remettent en cause la fiabilité de deux méthodes sélectionnées selon des études comparatives de la litérature, dans un cas d'utilisation différent, c'est-à-dire celui des vues synthétisées par DIBR. Cette sélection a été motivée par leur fiabilité, leur précision, leur efficacité ainsi que leur facilité de mise en œuvre.

Abbrev.	**Dénomination complète**	**Ref.**
DSIS	Double Stimulus Impairment Scale	[BT.93]
DSCQS	Double Stimulus Continuous Quality Scale	[BT.93]
SSNCS	Single Stimulus Numerical Categorical Scale	[BT.93]
SSCQE	Single Stimulus Continuous Quality Evaluation	[BT.93]
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation	[BT.93]
ACR	Absolute Category Rating	[ITU08]
ACR-HR	Absolute Category Rating with Hidden Reference removal	[ITU08]
DCR	Degradation Category Rating	[ITU08]
PC	Pair Comparison	[ITU08]
SAMVIQ	Subjective assessment Methodology for Video Quality	[ITU08]

Table 1.1 – *Méthodes d'évaluation subjective de la qualité pour les media 2D.*

Brotherton *et al.*[BHHB06] ont étudié la pertinence des méthodes ACR et SAMVIQ pour l'évaluation des méthodes 2D. Les résultats ont montré que la méthode ACR permet de présenter plus de stimuli que la méthode SAMVIQ. Dans l'étude [HGS+11], les résultats ont montré que la méthode ACR rend des scores fiables, peu importe l'échelle (5 points discrets, 5 points continus, 11 points discrets, onze points continus).

Dans notre étude, nous considérons les méthodes ACR-HR et Paired Comparisons (PC) pour leur précision.

Absolute Categorical Rating with Hidden Reference Removal (ACR-HR). Cette méthode consiste à présenter un seul stimulus à la fois aux observateurs. Les objets sont notés indépendamment selon une échelle de catégories. L'image de référence de chaque

5	Excellent
4	Bon
3	Moyen
2	Médiocre
1	Mauvais

Table 1.2 – *Echelle des catégoris pour la méthode ACR-HR*

stimulus est inclue dans la série de stimuli à noter, mais l'observateur n'en a pas connaissance. D'oú le terme anglais de *hidden reference* (référence cachée litéralement). A partir des scores obtenus par tous les observateurs, on peut calculer le score moyen MOS (mean opinion score) et le score différentiel DMOS (Differential Mean Opinion Score qui est une différence entre le MOS pour un stimulus et le score obtenu pour sa référence). L'échelle à 5 catégories recommandée par ITU est représentée dans le Tableau 1.2.

La méthode ACR requiert un nombre suffisant d'observateurs pour minimiser l'effet contextuel (les stimuli précédemment visionnés influencent le jugement de la qualité des stimuli suivants, l'ordre de présentation des stimuli influence le jugement des observateurs). La précision des résultats augmente avec le nombre de participants.

Paired Comparisons (PC). Dans cette méthode, les stimuli sont présentés par paires à l'observateur. L'observateur sélectionne le stimulus de la paire qui satisfait le mieux le critère de jugement demandé (par l'exemple la meilleure qualité d'image). Les résultats des comparaisons par paires sont stockés dans une matrice : chaque élément correspond à la fréquence à laquelle un stimulus est préféré à un autre. Ces données sont ensuite converties vers une échelle de valeurs en utilisant le modèle Thurstone-Mosteller ou le modèle Bradley-Terry's [Han01]. On obtient un continuum perceptuel hypothétique de valeurs de MOS.
Dans nos expériences, on utilise le modèle Thurstone-Mosteller et on demande aux observateurs de choisir le stimulus qu'il préfèrent dans chaque paire présentée. Bien que cette méthode soit reconnue pour sa précision, elle est chronophage puisque le nombre de comparaisons à effectuer augmente avec le nombre de stimuli à évaluer.

Les métriques objectives utilisées dans ces expériences ont été choisies pour leur popularité et pour leur dispponibilité. Elles sont répertoriées dans le Tableau 1.3 avec une croix. Les métriques choisies comportent à la fois des méthodes strictement basées "signal", des méthodes basées sur des principes de la perception humaine des images, des méthodes basées sur l'analyse de la structure des images, des méthodes reposant sur des modélisations du système visuel humain (HVS pour *human visual system*). La Figure 1.2 illustre cette classification. Les expériences en condition monoscopique ont été menées avec 43 observateurs pour les images fixes et 32 pour les séquences vidéo. Le Tableau 1.4 résume les informations relatives aux expériences en condition monoscopique. On désigne par *key frames* les images fixes arbitrairement choisies dans la séquence vidéo. Les expériences en condition stéréoscopique ont été menées avec 25 observateurs. Le Tableau 1.5 résume les informations relatives aux expériences en condition stéréoscopique.

	Objective metric	Abbrev.	Tested
Signal	Peak Signal to Noise Ratio	PSNR	X
Perception humaine	Universal Quality Index	UQI	X
	Information Fidelity Criterion	IFC	X
	Video Quality Metric	VQM	X
	Perceptual Video Quality Measure	PVQM	
Structure	Single-scale Structural SIMilarity	SSIM	X
	Multi-scale SSIM	MSSIM	X
	Video Structural Similarity Measure	V-SSIM	X
	Motion-based Video Integrity Evaluation	MOVIE	
HVS	PSNR- Human Visual System	PSNR-HVS	X
	PSNR-Human Visual System Masking model	PSNR-HVSM	X
	Visual Signal to Noise Ratio	VSNR	X
	Weighted Signal to Noise Ratio	WSNR	X
	Visual Information Fidelity	VIF	X
	Noise Quality Measure	NQM	X
	Moving Pictures Quality Metric	MPQM	

Table 1.3 – *Liste de méthodes d'évaluation de qualité objective d'images et de vidéos couramment utilisées.*

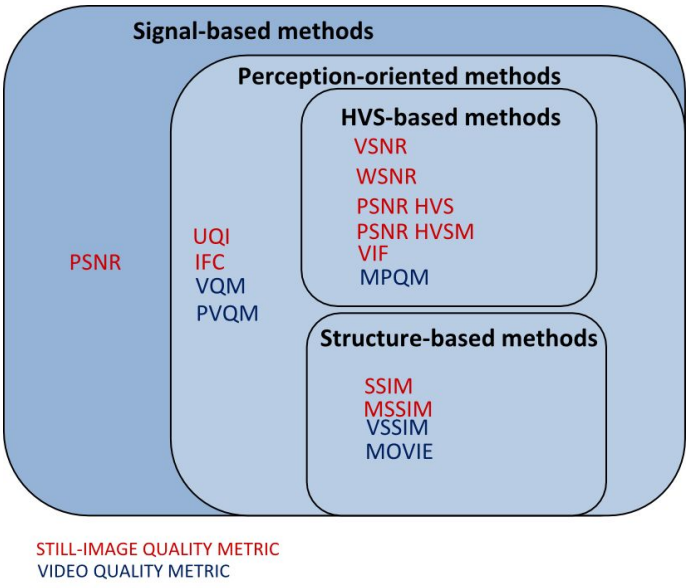


Figure 1.2 – *Classification possible des métriques de qualité selon la proposition de [PÓ8].*

		Expérience 1(images fixes)	**Expérience 2 (séquences vidéo)**
Stimuli		*Key frames* de chaque vue synthétisée	Séquence de la vue synthétisée
Tests subjectifs	**Nb. de participants**	43	32
	Méthodes	ACR-HR, PC	ACR-HR
Mesures objectives		Toutes les métriques disponibles de MetriX MuX	VQM, VSSIM, ET les métiques d'images fixes

Table 1.4 – *Présentation des expériences en condition monoscopique.*

		Expérience 3(images fixes stéréoscopiques)
Stimuli		Les paires stéréoscopiques sont composées des *key frames* de chaque vue synthétisée (à droite ou à gauche) , et de la vue originale droite ou gauche correspondante
Tests subjectifs	**Nb. de participants**	25
	Méthodes	ACR-HR
Mesures objectives		Toutes les métriques disponibles de MetriX MuX

Table 1.5 – *Présentation des expériences en condition stéréoscopique.*

Résultats

Les résultats obtenus en condition monoscopique avec les images fixes sont illustrés dans le Tableau 1.6 et ceux obtenus avec les séquences vidéo sont répertoriés dans le Tableau 1.7.

Dans le tableau 1.6, on donne les scores obtenus à partir des méthodes d'évaluation de qualité objective et subjective ainsi que le classement des algorithmes DIBR selon les scores de qualité obtenus. Dans le cas de la vidéo comme des images fixes, on constate que les métriques objectives sont cohérentes entre elles. Les classements obtenus par les méthodes d'évaluation de la qualité subjective (DMOS et PC) - dans le cas des images fixes - sont également cohérents entre eux. En revanche, on remarque que les classements issus des méthodes d'évaluation de la qualité subjective sont sensiblement différents des classements obtenus à partir des métriques objectives. En particulier, l'algorithme $A1$ est classé par les métriques objectives comme étant l'algorithme donnant les pires dégradations. Pourtant les méthodes d'évaluation de qualité subjective le classent comme étant celui donnant les dégradations les moins gênantes. Ceci peut s'expliquer par le fait que l'algorithme $A1$ réalise la syntèse de la vue en coupant les bords de l'image puis en interpolant l'image pour atteindre la taille originale. Ceci induit des déplacements d'objets, des changement de taille d'objets. Bien que les objets de la scène semblent de qualité correcte aux observateurs, les métriques objectives, basées sur le signal, pénalisent ces déplacements d'objets.

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.572	3.308	3.145	3.401	3.496	3.320	2.277
Classement	1	5	6	3	2	4	7
PC	1.776	0.779	0.338	0.825	1.745	0.610	-2.943
Classement	1	4	6	3	2	5	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Classement	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Classement	7	4	6	1	2	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Classement	7	4	6	1	2	3	5
VSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Classement	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Classement	7	5	6	2	3	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Classement	7	5	6	2	3	1	4
UQI	0.352	0.672	0.589	0.606	0.605	0.606	0.673
Classement	7	2	6	3	5	4	1
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Classement	7	5	6	2	3	1	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Classement	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Classement	7	4	5	3	2	1	6
PSNR HSVM	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Classement	7	4	5	3	2	1	6
PSNR HSV	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Classement	7	4	5	3	2	1	6

Table 1.6 – *Classement des algorithmes selon les mesures de qualité (images fixes).*

Les résultats obtenus en condition stéréoscopique avec les images fixes sont illustrés dans le Tableau 1.8. La remarque essentielle qui découle de l'analyse de ces résultats concerne la différence de classement des algorithmes par les méthodes d'évaluation de la qualité subjective entre les conditions monoscopique et stéréoscopique. En particulier, les algorithmes les moins bien classés par ces méthodes en monoscopique, se retrouvent mieux

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.523	3.237	2.966	2.865	2.789	2.956	2.104
Classement	1	2	3	5	6	4	7
PSNR	19.02	24.99	23.227	25.994	26.035	26.04	20.89
Classement	7	4	5	3	2	1	6
SSIM	0.648	0.844	0.786	0.859	0.859	0.859	0.824
Classement	7	4	6	1	1	1	5
MSSIM	0.664	0.932	0.825	0.948	0.948	0.948	0.888
Classement	7	4	6	1	1	1	5
VSNR	13.14	20.41	18.75	21.786	21.965	21.968	20.73
Classement	7	5	6	3	2	1	4
VIF	0.129	0.393	0.313	0.423	0.423	0.424	0.396
Classement	7	5	6	2	2	1	4
VIFP	0.153	0.415	0.342	0.446	0.446	0.446	0.419
Classement	7	5	6	1	1	1	4
UQI	0.359	0.664	0.58	0.598	0.598	0.598	0.667
Classement	7	5	6	3	3	3	1
IFC	0.779	2.399	1.926	2.562	2.562	2.564	2.404
Classement	7	5	6	2	2	1	4
NQM	8.66	15.933	13.415	16.635	16.739	16.739	10.63
Classement	7	4	5	3	1	1	6
WSNR	14.41	20.85	18.853	21.76	21.839	21.844	16.46
Classement	7	4	5	3	2	1	6
PSNR HSVM	13.99	19.37	18.361	21.278	21.318	21.326	16.23
Classement	7	4	5	3	2	1	6
PSNR HSV	13.74	19.52	17.958	20.795	20.823	20.833	15.91
Classement	7	4	5	3	2	1	6
VSSIM	0.662	0.879	0.809	0.899	0.898	0.893	0.854
Classement	7	4	6	1	2	3	5
VQM	0.888	0.623	0.581	0.572	0.556	0.557	0.652
Classement	7	5	4	3	1	2	6

Table 1.7 – *Classement des algorithmes selon les mesures de qualité (séquences vidéo)*

classés en stéréoscopique. On suppose que les artefacts gênants en monoscopique peuvent être masqués en stéréoscopique. Une autre hypothèse possible concerne les informations stéréoscopiques contradictoires qui peuvent causer une gêne visuelle et conduire à des scores plus faibles. En effet, dans le cas de l'algorithme $A1$, les vues synthétisées impliquant un déplacement relatif ou un changement de taille relatif des objets, les correspondances stéréoscopiques sont difficilement gérées par le système visuel humain.

Des analyses supplémentaires ont été menées pour estimer la fiabilité des méthodes utilisées (pour la qualité subjective et la qualité objective). Les tableaux 1.9, 1.10 et 1.11 concernent l'étude de la fiabilité des méthodes d'évaluation de la qualité subjective. Les tableaux 1.12, 1.13 et 1.14 concernent l'étude de la fiabilité des méthodes d'évaluation de la qualité objective.

Dans les tableaux 1.9, 1.10 et 1.11, les nombres entre parenthèse indiquent le nombre minimum d'obsevateurs requis pour atteindre une distinction statistique entre deux distributions données (en sachant que VQEG recommande un minimum de 24 participants dans le Multimedia Test Plan [Gro], on note en gras les valeurs supérieurs à 24) . Ces tableaux montrent que dans la plupart des cas, plus de 24 observateurs sont requis pour obtenir une distinction statistique entre deux distributions (relatives à deux algorithmes de synthèse).

Dans les tableaux 1.12, 1.13 et 1.14 les coefficients de corrélation de Pearson entre les scores subjectifs et les scores objectifs sont présentés. On constate que les métriques objectives ne sont pas corrélées avec les notes subjectives. Ces résultats suggèrent que les métriques objectives ne sont pas adaptées à l'évaluation des media incluant des artefacts liés au pro-

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.647	3.637	3.660	3.678	3.658	3.662	3.548
Classement	5	6	3	1	4	2	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Classement	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Classement	7	4	6	1	2	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Classement	7	4	6	1	2	3	5
VSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Classement	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Classement	7	5	6	2	3	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Classement	7	5	6	2	3	1	4
UQI	0.352	0.672	0.589	0.606	0.605	0.606	0.673
Classement	7	2	6	3	5	4	1
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Classement	7	5	6	2	3	1	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Classement	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Classement	7	4	5	3	2	1	6
PSNR HSVM	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Classement	7	4	5	3	2	1	6
PSNR HSV	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Classement	7	4	5	3	2	1	6

Table 1.8 – *Classement des algorithmes selon les mesures de qualité (images fixes stéréoscopiques).*

cessus de synthése DIBR. La difficulté vient du fait que les sources de dégradation des vues synthétisées sont d'une part multiples (dans le sens où les types d'artefacts varient selon les stratégies) et localisées le long des bords des objets de la scène.

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑**(32)**	↑(<24)	↑**(32)**	**o (>43)**	↑**(30)**	↑(<24)
$A2$	↓**(32)**		↑(<24)	**o (>43)**	**o (>43)**	**o (>43)**	↑(<24)
$A3$	↓(<24)	↓(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓**(32)**	**o(>43)**	↑(<24)		**o(>43)**	**o(>43)**	↑(<24)
$A5$	**o(>43)**	**o(>43)**	↑(<24)	**o(>43)**		↑**(28)**	↑(<24)
$A6$	↓**(30)**	**o(>43)**	↑(<24)	**o (>43)**	↓**(28)**		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓ (<24)	↓(<24)	↓(<24)	

Table 1.9 – *Résultats du test de Student avec les résultats de l'ACR-HR (images fixes). Légende :↑ : supérieur, ↓ : inférieur, o : statistiquement équivalent. Lecture : Ligne"1" est statistiquement supérieur à colonne "2". La distinction est stable quand "32" observateurs participent.*

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)
$A2$	↓(<24)		↑**(28)**	o(<24)	↓(<24)	**o(>43)**	↑(<24)
$A3$	↓(<24)	↓**(28)**		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓(<24)	**o(>43)**	↑(<24)		↓(<24)	↑**(43)**	↑(<24)
$A5$	↓(<24)	↑(<24)	↑(<24)	↑(<24)		↑(<24)	↑(<24)
$A6$	↓(<24)	**o(>43)**	↑(<24)	↓**(<43)**	↓(<24)		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	

Table 1.10 – *Résultats du test de Student avec les résultats des ?Paired comparisons? (images fixes). Légende :↑ : supérieur, ↓ : inférieur, o : statistiquement équivalent. Lecture : Ligne"1" est statistiquement supérieur à colonne "2". La distinction est stable quand "32" observateurs participent.*

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑(7)	↑(3)	↑(3)	**2**	↑(3)	↑(1)
$A2$	↓(7)		↑(2)	↑(2)	↑(1)	↑(2)	↑(1)
$A3$	↓(3)	↓(2)		**o(>32)**	↑(9)	**o(>32)**	↑(1)
$A4$	↓(3)	↓(2)	**o(>32)**		**o(>32)**	**o(>32)**	↑(1)
$A5$	↓(2)	↓(1)	↓(9)	**o(>32)**		↑(15)	↑(1)
$A6$	↓(3)	↓(2)	**o(>32)**	**o(>32)**	↑(15)		↑(1)
$A7$	↓(1)	↓(1)	↓(1)	↓ (1)	↓(1)	↓(1)	

Table 1.11 – *Résultats du test de Student avec les résultats de l'ACR-HR (séquences video). Légende :↑ : supérieur, ↓ : inférieur, o : statistiquement équivalent. Lecture : Ligne"1" est statistiquement inférieure á la colonne"2". La distinction est stable quand "32" observateurs participent.*

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**	**UQI**	**IFC**	**NQM**	**WSNR**	$PSNR_{HVSM}$	$PSNR_{HVS}$
CC_{DMOS}	50.0	40.4	57.4	35.0	31.3	22.2	19.1	22.3	57.2	47.7	44.3	42.7
CC_{PC}	36.4	23.1	43.2	16.8	18.2	18.3	24.8	17.7	37.9	33.9	37.5	36.6

Table 1.12 – *Coefficients de corrélation de Pearson entre les scores DMOS et les scores objectifs en pourcentage (images fixes).*

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**	**UQI**
CC_{DMOS}	33.80	40.84	49.95	38.87	27.01	20.00	28.70
	IFC	**NQM**	**WSNR**	$PSNR_{HVSM}$	$PSNR_{HVS}$	**VSSIM**	**VQM**
CC_{DMOS}	21.68	41.15	27.98	31.47	29.76	39.37	43.41

Table 1.13 – *Coefficients de corrélation de Pearson entre les scores DMOS et les scores objectifs en pourcentage (séquences video).*

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**
CC_{DMOS}	46.98	45.06	60.86	26.44	38.46	42.96
	UQI	**IFC**	**NQM**	**WSNR**	$PSNR_{HVSM}$	$PSNR_{HVS}$
CC_{DMOS}	31.72	40.96	52.66	51.58	46.59	46.13

Table 1.14 – *Coefficients de corrélation de Pearson entre les scores DMOS et les scores objectifs en pourcentage (images fixes stéréoscopiques).*

Conclusion

Les études présentées dans cette sous-section cherchaient à déterminer la pertinence de l'utilisation des méthodes d'évaluation de qualité d'images/vidéo, généralement employées en imagerie 2D, pour le cas spécifique des vues synthétisées. Les résultats ont montrés que les métriques objectives de qualité ne parviennent pas à prédire de façcon fiable la qualité perçue par des observateurs. Il s'est également avéré que les artéfacts non gênants en visualisation monoscopique pouvaient l'être en visualisation stéréoscopique. Par ailleurs, le nombre de participants aux tests d'évaluation subjective de qualité s'est révélé supérieur aux recommandations de VQEG, d'après nos analyses statistiques.

1.2 Compression des cartes de profondeur basée sur la méthode LAR

En l'absence de méthode de compression normalisée pour les données MVD, on se propose dans ces travaux de thèse d'adresser le problème de la compression des cartes de profondeur. La plupart des méthodes proposées dans la litérature reposent sur des extensions de codecs de l'état de l'art pour les images ou vidéos 2D. La plus populaire est

H264/AVC [STL04] dont l'extension 3D (H.264/MVC pour Multi-view Video Coding) pour les données MVV a fait l'objet de nombreuses adaptations [MMSW06] pour la compression des données MVD [MSD+09] sans réaliser un gain suffisant. Actuellement, l'organisme de standardisation MPEG considére la normalisation d'un nouveau standard pour les MVD : 3DVC. De récentes études ont montré que les méthodes de codage couramment utilisées pour les medias 2D, appliquées aux cartes de profondeur, induisent des dégradations gênantes à l'issue de la synthèse de vue. En effet la difficulté de conception d'une méthode robuste réside essentiellement dans le fait que l'impact des dégradations de compression sur les données de profondeur peut être fatal à la qualité de la vue synthétisée, comme cela a été montré dans de récentes études [MMS+09]. Dans ces travaux de thèse, nous proposons deux extensions d'une méthode de compression dont l'efficacité a été prouvée pour le cas des images fixes conventionnelles. Nous y apportons des modifications permettant son adaptation à l'encodage des cartes de profondeur. les deux méthodes proposées ont été baptisées *Z-LAR* et *Z-LAR-RP*. Elles sont nées de l'étude méticuleuse des performances de la méthode de base pour l'encodage des cartes de profondeur.

Les deux méthodes proposées sont basées sur une représentation de l'image par un quad-tree noté **Quad-tree**$^{[N_{max}\dots N_{min}]}$. La décomposition du quad-tree repose sur un critére de d'homogénéité. Soit **Quad-tree**$^{[N_{max}\dots N_{min}]}$ le quad-tree, où N_{max} et N_{min} sont les tailles maximum et minimum permises pour les blocs du quad-tree, respectivement. Soit $I(x,y)$ le pixel de coordonnées (x,y) dans l'image I et $I(b^N(i,j))$ est le bloc $b^N(i,j)$ dans l'image I, comme décrit ci-dessous :

$$b^N(i,j) = \{(x,y) \in N_x \times N_y \mid N \times (i+1),\ \text{et}\ N \times j \leq y \leq N \times (j+1)\} \tag{1.1}$$

La décomposition du quad-tree repose sur la détection de l'activité locale. On considère un support, la différence entre la valeur de luminance maximale et la valeur de luminance minimale de ce support est calculée. Pour une partition donnée **Quad-tree**$^{[N_{max}\dots N_{min}]}$ de l'image I, pour tout pixel $I(x,y)$, la taille du bloc auquel un pixel appartient est exprimée par :

$$Size(x,y) = \begin{cases} N \in [N_{max} \dots N_{min}[\quad \text{if} \\ \quad |max\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) - min\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) \leq Y \\ \qquad\qquad \text{et si } \exists (k,m) \in \{0,1\}^2 \\ |max\big(I(b^N(\lfloor \frac{x+k}{N/2} \rfloor, \lfloor \frac{y+m}{N/2} \rfloor))\big) - min\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) > Y \\ N_{min} \qquad\qquad\qquad\qquad \text{sinon.} \end{cases} \tag{1.2}$$

où $min\big(I(b^N(i,j))\big)$ et $max\big(I(b^N(i,j))\big)$ sont les valeurs minimales et maximales du bloc $I(b^N(i,j))$ respectivement, et Y est le seuil d'homogénéité. La valeur du seuil utilisée pour réaliser la décomposition du quad-tree détermine la représentation finale de l'image. Dans la méthode LAR, l'image étant considérée comme composée d'une image de basse résolution et d'une image contenant les détails, l'image de basse résolution de l'image I (i.e., notée l'image flat) est obtenue en affectant à chaque bloc du quad-tree la valeur moyenne de ses pixels. Soit LR l'image de basse résolution, chacun de ses pixels $LR(x,y)$ est défini par :

$$LR(x,y) = \frac{l}{N^2} \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} I(\lfloor \frac{x}{N} \rfloor \times N + k, \lfloor \frac{y}{N} \rfloor \times N + m) \tag{1.3}$$

où $N = Size(x,y)$. Le schéma de base de la méthode LAR a donné suite à de nombreux travaux d'extension parmi lesquels le profil pyramidal. Il a été conu pour augmenter les

Algorithm 1.1: Prediction of lowest level of the pyramidal decomposition

Require: $\tilde{L}_l$ is the estimated representation of the image at the decoder side, for level l,
Quad-tree$^{[N_{max}...N_{min}]}$ is the quad-tree partition.
for $l = l_{max} \dots l_1$ **do**
Estimate $\tilde{L}_l$ as in the LAR method
end for
for each block of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N = N_{max} \dots N_{min}$ **do**
Given **Quad-tree**$^{[N_{max}...N_{min}]}$, then $\tilde{L}_0(b^N(i,j)) = \tilde{L}_1$
end for
for each block of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N = N_{min}$ **do**
$\tilde{L}_0(b^(N_{min})(i,j))$ = Mean value of the closest block b^N of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N > N_{min}$
end for
return $\tilde{L}_0$

capacités de scalabilité et adresser la compression sans pertes. On connaît ses extensions sous le nom de Interleaved S+P [BDR05, PBD+08] et RWHaT+P [DBBC08]. Dans la suite, nous utiliserons le profil dit Interleaved S+P.

Les deux méthodes que nous proposons dans ces travaux remettent en cause la stratégie de distribution du débit et sont basées sur la même approche pour adresser le problème. En effet, dans les travaux précédents de Pasteau *et al.* [PBD+10], la stratégie recommandée consiste à appliquer un pas de quantification dépendant de la taille des blocs du quad-tree, pour les images conventionnelles. Bien que cette méthode ait montré son efficacité pour les images conventionnelles, nos études ont montré que cette approche n'est pas adaptée à l'encodage des cartes de profondeur. En effet, cette méthode, basée sur des observations des propriétés du système visuel humain, inflige des quantifications grossières aux plus petits blocs. Or les petits blocs de la carte de profondeur correspondent aux frontières de discontinuité très importantes pour la phase de synthèse de point de vue virtuel. Pour cette raison, nous proposons de modifier cette stratégie pour privilégier une représentation de l'image variable en fonction du débit cible. En effet, en augmentant la valeur du seuil d'homogénéité Y, on peut réaliser un gain de débit tout en conservant les éléments essentiels la structure de la carte de profondeur.

Z-LAR : une nouvelle méthode pour l'encodage des cartes de profondeur

L'approche dite *Z-LAR* propose d'utiliser le profil pyramidal de la méthode LAR et d'encoder la pyramide jusqu'au niveau directement supérieur à la pleine résolution. L'Algorithme 1.1 rappelle les étapes fondamentales de l'approche pour atteindre la pleine résolution. Un filtre est ensuite appliqué pour éliminer l'effet de bloc au niveau des plus petits blocs décodés. Ce filtre multilatéral prend en compte l'information de la couleur décodée associée à la profondeur. la couleur peut être encodée via toute méthode d'encodage d'images de l'état de l'art. Le filtre utilisé est défini par l'équation Eq. 1.5. L'image de couleur décodée est notée $\tilde{C}$ L'image filtrée est notée $\tilde{L}_{0r}$ et chaque pixel est noté $\tilde{L}_{0r}(x,y)$. On définit également Ω l'ensemble de pixels tels que :

$$\Omega = \left\{ \tilde{L}_0(x,y) \,\middle|\, \tilde{L}_0(x,y) \in \tilde{L}_0\left(b^N(i,j)\right), \quad N \in [N_{min} \dots 4] \right. \tag{1.4}$$

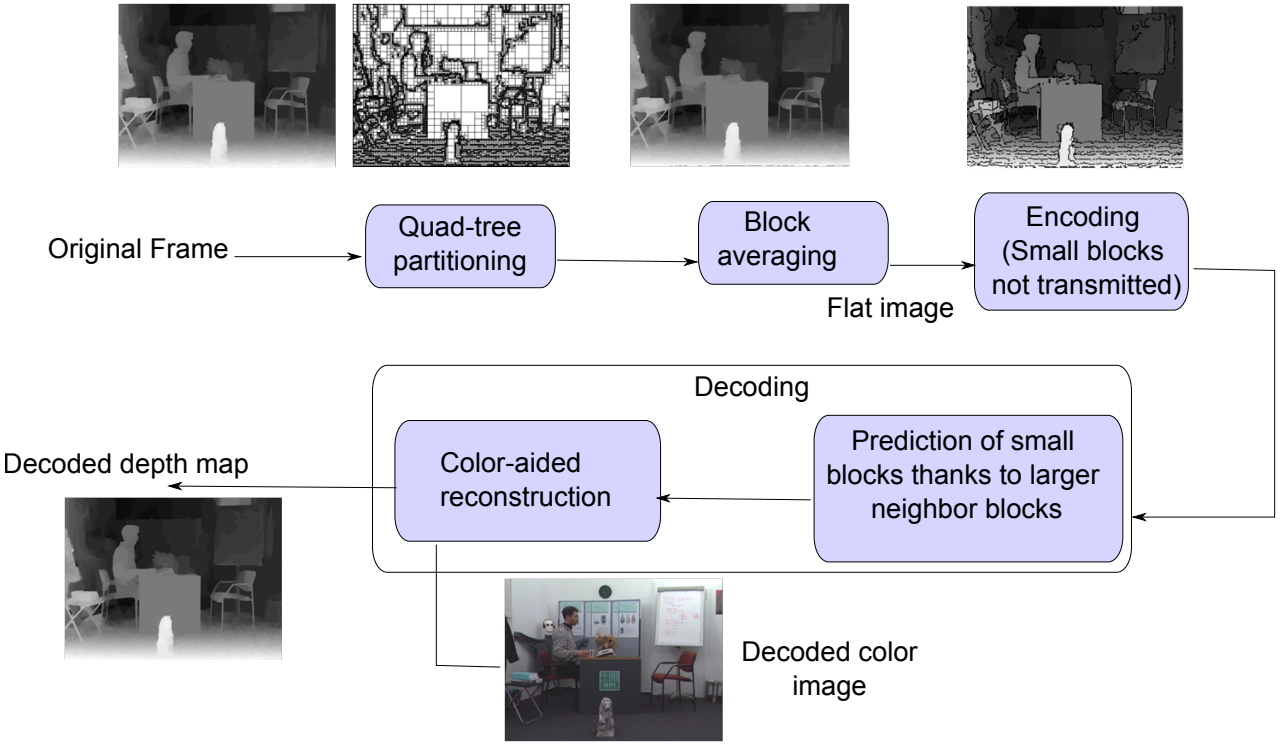


Figure 1.3 – *Schéma global de la schéma proposée.*

$\forall\ \tilde{L}_0\,(x,y) \in \Omega,$

$$\tilde{L}_{0r}\,(x,y) = \tilde{L}_{0r}\,(p) = \frac{1}{K}\sum_{q\in\Gamma} \tilde{L}_0\,(p) e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|\tilde{L}_0(p)-\tilde{L}_0(q)\|}{2\sigma_s}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}} \tag{1.5}$$

avec

$$K = \sum_{q\in\Gamma} e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|\tilde{L}_0(p)-\tilde{L}_0(q)\|}{2\sigma_s}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}} \tag{1.6}$$

où Γ est le support utilisé pour le calcul ; $Luma$ est la luminance de la couleur décodée ; $Luma(p)$ et $Luma(q)$ sont les pixels de luminance de la couleur décodée ; σ_d, σ_s, σ_c sont les écart-type relatifs au domaine spatial, au domaine des valeurs de la profondeur (similarité des valeurs de profondeur), et le domaine des valeurs de la couleur, respectivement. La figure 1.3 donne un schéma global de la méhode proposée. Dans cette figure, à l'étape d'encodage, les blocs noirs correspondent aux blocs non-encodés.

La méthode a été comparée à la méthode H.264/AVC en intra et a donné de meilleurs résultats en termes de qualité visuelle. Les métriques objectives (PSNR et VIF) ont donné des résultats contradictoires puisque le PSNR juge la méthode proposée moins acceptable que la méthode H.264 ; alors que le VIF donne la méthode proposée comme étant meilleure. En particulier cette étude a prouvé que notre méthode élimine les artefacts de crénelage le long des discontinuités des objets.

Z-LAR-RP : la prédiction hiérarchique basée région dans Z-LAR

Dans cette nouvelle approche, nous cherchons à la fois à permettre la multirésolution en tirant parti de l'encodage des niveaux de la pyramide et à augmenter les performances en termes de qualité de vue synthétisée et de gain en simplicité de traitement. On peut donc, avec la méthode proposée, encoder des niveaux de la pyramide LAR supérieurs à la pleine résolution et réaliser des économies de débit. La taille originelle de l'image est atteinte grâce à une méthode basée sur la connaissance des régions de l'image. La figure 1.4 donne un schéma global de la méthode employée.

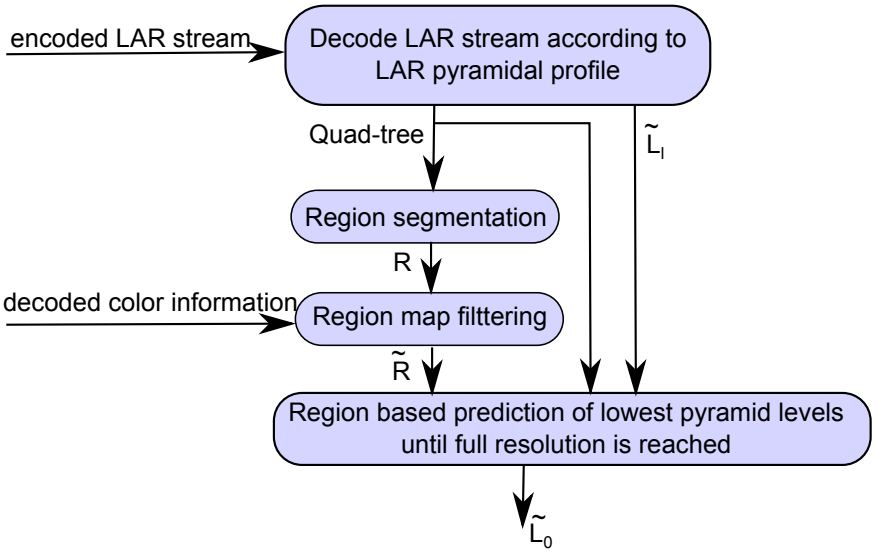


Figure 1.4 – *Schéma global de la schéma proposée.*

La méthode de segmentation de régions utilisée dans cette approche se base sur les travaux de Strauss [Str11]. Elle est appliquée au quad-tree décodé de la profondeur. La particularité de cet algorithme tient du fait qu'il ne requiert que l'information structurelle de l'image, qui est contenue dans le quad-tree. Or ce quad-tree est un composant du flux LAR encodé. Au décodeur on peut donc décoder le quad-tree et appliquer cet algorithme pour obtenir une carte de régions R. Nous utilisons cette carte de régions pour décoder les niveaux successifs de la pyramide de la carte de profondeur. Nous réalisons ensuite un raffinement de cette carte de régions R en utilisant la couleur décodée associée à la carte de profondeur couramment décodée zfin de mettre en cohérence les contours de la carte de profondeur et ceux de l'image d écodée de couleur associée. Notons chaque pixel de R comme $R(i,j)$ ou bien comme $R(p)$. La carte de régions filtrée est notée $\tilde{R}$.
Pour chaque pixel, un support Γ_p est considéré, qui est une fenêtre de pixels centrée sur le pixel couramment traité. La composante de la luminance *Luma* de la couleur décodée est utilisée pour estimer la similarité des pixels voisins. L'algorithme 1.2 a pour but d'affecter à chaque pixel de la carte de régions l'étiquette de région la plus probable selon la région d'appartenance des voisins. La distance au pixel central et la similarité de la couleur des voisins sont pris en compte. Cette contrainte est exprimée par les facteurs σ_c and σ_d respectivement.

Les niveaux l_{max} à l_{min} de la pyramide sont encodés et transmis. Le niveau l_{min} peut être choisi librement entre 1 et $l_{max} - 1$. Le décodage de la pleine résolution de la carte de profondeur est possible grâce à la carte des régions filtrée. Tout pixel de coordonnées

Algorithm 1.2: Filtrage de la carte de régions en fonction de la couleur décodée

Require: R the region map with $N_{regions}$ labels ;
$W[N_{regions}]$ the array of region weights ;
$Luma$ the associated decoded texture image
Initializations
$Temp(p) = Temp(i,j) = R(p) = R(i,j) \mid \{(i,j) \in N_x \times N_y\}$
$W[k] = 0 | \{k \in [1 \dots N_{regions}]\}$
for all $p \in R$ **do**
  for all $q \in \Gamma_p$ **do**
    $r = R(q)$
    $W[r] = W[r] + e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}}$
  end for
  Find $\tilde{r} \mid \tilde{r} = \underset{k \in [1 \dots N_{regions}]}{\text{argmax}} \; W[k]$
  $Temp(p) = \tilde{r}$
  Reset tous les éléments de W à 0
end for
$\tilde{R}(i,j) = Temp(i,j) \mid \{(i,j) \in N_x \times N_y\}$
return $\tilde{R}$

(i,j) est noté p. $\tilde{L}_{l_{min}}$ le niveau le plus bas de la pyramide réllement encodé et transmis, avec $l_{min} \geq 1$, $l = 0$ étant la pleine résolution. Le bloc $b^N(i,j)$ est décrit par l'équation 1.1 et N est la taille du bloc définie par l'équation 1.2. Pour chaque pixel,un support Γ_p est considéré, qui est une fenêtre de pixels centrée sur le pixel couramment traité. K est un facteur de normalisation :

$$K = \sum_{q \in \Gamma_2} \delta_p(q) e^{-\frac{\|p-q\|}{2\sigma_1}} e^{-\frac{\|\tilde{L}_l(p)-\tilde{L}_l(q)\|}{2\sigma_2}}, \tag{1.7}$$

où $\delta_p(q)$ la fonction d'existance définie par :

$$\delta_p(q) = \begin{cases} 1 & \text{if } \tilde{R}(p) = \tilde{R}(q) \\ 0 & \text{sinon} \end{cases} \tag{1.8}$$

La reconstruction du niveau le plus bas est donc une somme pondérée de la valeur de profondeur des pixels voisins. Les pixels voisins ne contribuent à cette somme que s'ils appartiennent à la même étiquette de région dans l'image de pleine résolution.

1.2 Relations entre la couleur et la profondeur pour l'allocation de débit

Dans ces travaux de thèse, nous nous sommes également intéressés à la répartition du débit entre la texture et la profondeur lors de la compression de séquences MVD. Les effets de la quantification sur les deux types de données ont été étudiés. La distorsion est mesurée sur les images synthétisées à partir des séquences MVD encodées et décodées par la méthode H.264/MVC dans un premier cas d'étude, et HEVC dans un second cas. Nous avons considéré l'étude de deux méthodes de compression différentes dans le but de contrôler l'influence du choix de la stratégie d'encodage dans la répartition du débit optimisant la qualité de la vue synthétisée. Bien que la profondeur soit codée sur une seule composante (contre trois pour la texture), allouer 25% du débit à la profondeur n'est pas

Algorithm 1.3: Prédiction basée sur la carte des régions filtrée

Require: $\tilde{L_l}$ the lowest decoded level image of the pyramid ;
$\textbf{Quad-tree}^{[N_{max}...N_{min}]}$ the quad-tree partition ;
$\tilde{R}$ the filtered region map.
repeat
  for all $p \in \tilde{L}_{l-1}$ **do**
    if $\tilde{L}_{l-1} \in b^N \mid N < 2^l$ **then**
      $\tilde{L}_{l-1}(p) = \frac{1}{K} \sum_{q \in \Gamma_2} \delta_p(q) e^{-\frac{\|p-q\|}{2\sigma_1}} e^{-\frac{\|\tilde{L}_l(p)-\tilde{L}_l(q)\|}{2\sigma_2}}$
    else
      $\tilde{L}_{l-1}(p) = \tilde{L}_l(p)$
    end if
  end for
until $l = 0$

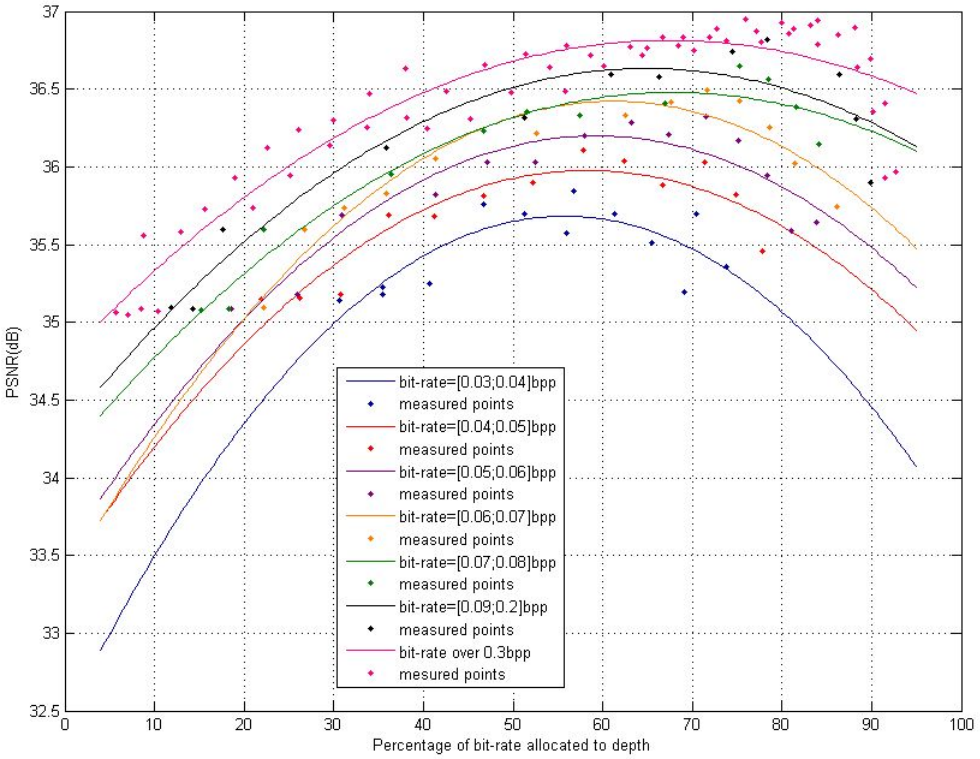
le choix optimal. Les résultats de nos expeérimentations ont montré que plus de la moitié du débit peut être réservée aux données de profondeur pour favoriser la qualité de la vue synthétisée (dans le cas de l'utilisation de H264/MVC). Dans le premier cas d' étude, en prenant en entrée des vidéos MVD, nous avons utilisé le codeur MVC pour compresser les textures d'un coté, et les cartes de profondeur de l'autre. Dans le deuxième cas, nous avons utilisé HEVC pour compresser les deux types de données également.
En faisant varier les pas de quantification, on observe l'évolution du PSNR lors de la synthèse de vue par VSRS, l'algorithme de syntése de vues de référence de MPEG(*View Synthesis Reference Software*[TFS+08]). La compression des textures fait apparaitre des zones floues, et la compression des cartes de profondeur entraîne des distorsions géométriques. A débit total fixé, nous avons constaté il existe un rapport idéal entre le débit alloué à la texture et celui alloué à la profondeur, maximisant la qualité visuelle de l'image générée. Pour assurer la meilleure qualité d'image lors de la synthèse de vue avec VSRS, nous avons montré qu'il est nécessaire d'attribuer aux cartes de profondeur de 30% à 60% du débit total en fonction de la méthode de compression utilisée. La figure 1.5 et la figure 1.6 illustrent les résultats obtenus dans le cas de l'utilisation de H.264/MVC et de HEVC respectivement.

Nos travaux ont mis en évidence les raisons des différences de distributions requises d'une séquence à l'autre. En particulier, la distance entre les caméra, et la surface des zones découvertes influencent le ratio nécessaire entre la couleur et la profondeur. L'entropie de l'information de profondeur joue également un rôle important : plus la structure de la scène est complexe, plus la part de la profondeur nécessaire pour optimiser la qualité de la synthèse sera importante.

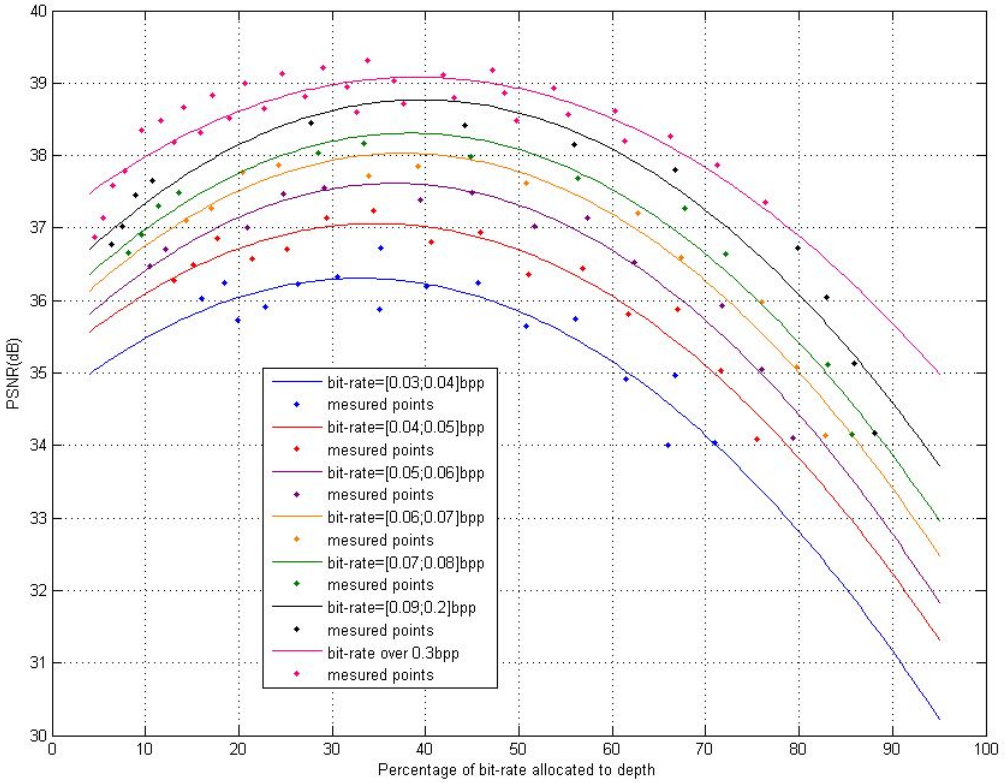
1.3 Conclusion

Cette thèse s'est intéressée aux questions de compression de données MVD et de qualité visuelle de vues reconstruites à partir d'informations décompressées ou non. Tout au long de ces travaux de thèse les préoccupations ont concerné la qualité visuelle, et l'optimisation des choix de codage s'est basée sur la perception humaine des vues synthétisées.

Des études ont été réalisées pour caractériser les artefacts liés aux algorithmes DIBR,

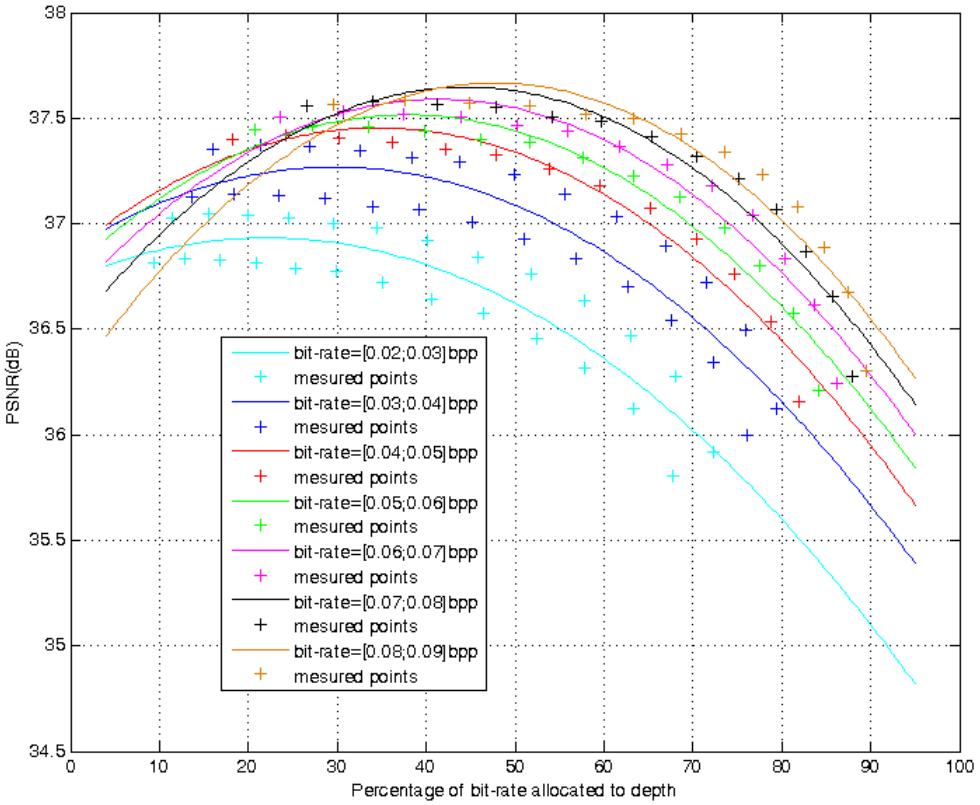


(a) PSNR (dB) des vues synthétisées en fonction du débit attribué à la profondeur en pourcentage par rapport au débit total pour Ballet

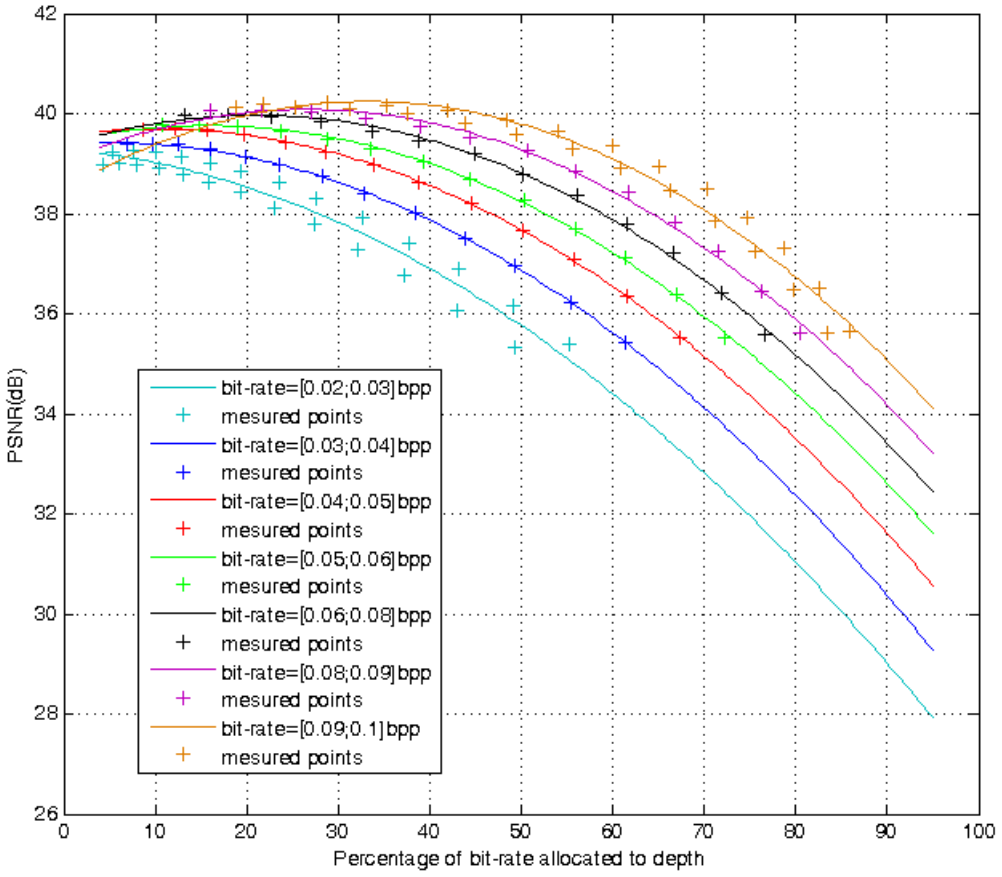


(b) PSNR (dB) des vues synthétisées en fonction du débit attribué à la profondeur en pourcentage par rapport au débit total pour Book Arrival

Figure 1.5 – *Courbes débit/distorsion interpolées des vues synthétisées, en utilisant des données encodées avec H.264/MVC.*



(a) PSNR (dB) des vues synthétisées en fonction du débit attribué à la profondeur en pourcentage par rapport au débit total pour Ballet



ainsi que pour déterminer l'adéquation des méthodes d'évaluation de qualité (subjective et objective) dans le cas des vues synthétisées par DIBR. Les expérimentations ont montré d'une part que les sources des dégradations sont multiples : la précision des séquences de profondeur (acquises ou estimées), le processus de synthèse lui-même (soumis à des arrondis de position des pixels), ainsi que les stratégies de remplissage (*inpainting*) choisies pour remplir les zones découvertes par le changement de point de vue, sont autant de facteurs influant sur la qualité d'une vue synthétisée.

Concernant l'évaluation de la qualité des vues synthétisées, le principe même du DIBR implique des cas d'utilisation (et donc d'évaluation) chevauchant à la fois les contextes 2D, conventionnels, et les contextes 3D (avec stéréovision). Les DIBR sont des méthodes utilisées en 3D pour réaliser des images 2D, mais ces images sont également utilisables pour des applications de stéréovision. Cette caractéristique particulière est probablement une des causes de la difficulté de mise en place de protocoles d'évaluation de qualité subjective ainsi que de techniques d'évaluation objective.

D'autre part, les résultats obtenus ont aussi montré que, dans un contexte d'utilisation en 2D, les techniques d'évaluation subjective et objective ne sont pas suffisamment adaptées pour évaluer correctement la qualité des vues synthétisées à partir d'algorithmes DIBR. Les mêmes vues synthétisées observées en condition de vue stéréoscopique n'ont pas abouti aux mêmes résultats. En particulier, les artefacts influant le moins la qualité des images en condition monoscopique se sont révélés gênants en condition stéréoscopique.

En ce qui concerne les études sur la compression des cartes de profondeur, nous avons proposé deux méthodes dérivées pour le codage des cartes de profondeur et basées sur un schéma de compression connu pour les images fixes conventionnelles. En nous appuyant sur nos observations, nous avons proposé une stratégie de représentation et de codage adaptée au besoin de préserver les discontinuités de la carte tout en réalisant des taux de compression importants. L'originalité des méthodes que nous proposons réside dans le fait que, grâce à la quantification spatiale que nous effectuons, lors de faibles débits requis, la carte de profondeur décodée est uniforme. En d'autres termes, pour les faibles débits, nous privilégions la qualité des contours des objets rendus au prix d'une profondeur de scène réduite. Nous renonçons aux choix de quantification grossière le long des discontinuités de profondeur, qui, bien que permettant des économies drastiques sur le débit, induisent des artefacts visuellement très gênants le long des bords des objets rendus.

Nous avons également réalisé des études sur la répartition du débit entre la texture et la profondeur lors de la compression de séquences MVD. Les résultats de nos expériences ont montré que cette répartition dépend de plusieurs facteurs. La méthode de compression utilisée pour encoder les données avant la synthèse de vue semble influencer la répartition du débit nécessaire pour une bonne qualité visuelle. Les propriétés du contenu (entropie des données, importance des zones découvertes) jouent également un rôle dans cette répartition.

Les résultats de cette thèse peuvent être utilisés pour aider à la conception de nouveaux protocoles d'évaluation de qualité de données de synthése, dans le contexte des données MVD ; pour la conception de nouvelles métriques de qualité ; pour améliorer les schémas de codage pour les données MVD, notamment grâce aux approches originales proposées ; pour optimiser les schémas de codage de données MVD, à partir de nos études sur les rela-

tions entre la texture et la profondeur. Nous espérons, dans l'avenir, susciter de nouvelles contributions s'intéressant à ces problématiques.

CHAPTER 1

Introduction

1.1 Objectives of the thesis and Contributions

Despite its long history, three dimensional video has gained a growing interest in research activities for the last decade. Improvement of hardware solutions has enabled the progress of three-dimensional (3D) video technologies. However, the success of the two main applications referred to as "3D Video" (namely 3D Television (3D TV) that provides depth to the scene, and Free Viewpoint Video (FVV) that enables interactive navigation inside the scene) relies on their ability to provide an added value (depth or immersion) coupled with high-quality visual content.
3D video requires the acquisition of multiple video streams. 3D scene representations such as Multi View Video (MVV) data, that consists of a set of conventional video streams, provide video sequences of the same scene at slightly different viewpoints. When associated to depth video streams, the scene representation is called Multi View plus Depth (MVD) data. Efficient compression schemes are expected to handle this huge amount of data.

All along the processing chain of 3D video, artifacts may be induced. In particular, the essential Depth-Image-Based-Rendering (DIBR) techniques in 3D Video, used for the generation of new virtual viewpoints, induce new types of artifacts. Since 3D Video success depends on the ability to provide high quality contents, determining the influence of the different sources of distortion in the synthesized views and their combined effect is primordial. This raises the issue of the deployed means addressing the assessment of three-dimensional media. Many studies already tackled the problem of compression of 3D video data, but few of them focused on the perceptual quality of the rendered virtual views as a mainstay of their coding strategies. Most of the proposed coding methods are based on 2D codecs. However, their efficiency regarding depth map compression is uncertain since they are generally optimized for enhancing human perception of color. Yet, depth maps are not natural images and erroneous depth values can lead to annoying distortions in the synthesized views.

Our research is supported by the PERSEE project, a French national research project (ANR), whose scientific work is in the direction of a content-based and perceptually driven

representation and coding paradigm using a clever combination of perceptual models and a rate-visual quality optimization framework among others.
This thesis is meant to investigate new MVD coding frameworks limiting as much as possible the perceptible distortions occurring in the views synthesized from decompressed data. Our objectives thus particularly concern the development of perceptually driven tools in the context of MVD coding. This is quite trendy since Ndjiki-Nya *et al.* also recently directed their efforts for proposing a 2D perceptual oriented video coding method in [NDK+12]. In our case, the difficulties we tackle lie in the fact that not only the distortions sources are multiple but there is no dedicated quality assessment tool for this specific type of data. Our research activities focused on the issues mentioned below and led to the following scientific contributions:

- *View synthesis related artifacts and Assessment of synthesized views* We conducted extensive tests over views synthesized from several synthesis algorithms. Each synthesis strategy induces specific types of artifacts that are evaluated through subjective evaluation tests and through objective quality metrics.

- *Impact of depth quantization and Design of a perception-oriented depth compression scheme* Our analyses of depth quantization through different coding methods helped in the understanding of the essential requirements for a depth compression scheme. Based on our observations, we proposed depth compression method adapted to the need for an edge-preserving method with efficient compression ratio. The originality and the distinctiveness of our proposed method lies in the fact that due to the spatial quantization, as the bit rate decreases, the reconstructed depth map tends to be flat. In other words, at low bit-rate, we give priority to edges quality with less depth feeling instead of coarse quantization around the edges leading to projection errors in the synthesized views, or visual discomfort in stereoscopic conditions. Two different encoding methods are proposed.

- *Bit-rate allocation between texture and depth data and the influencing parameters* We have conducted studies on the question of bit-rate allocation between texture and depth data, in the context of MVD compression. The analysis of the relationships between texture and depth data in the context of bit allocation is essential because they contribute to the visual quality of the synthesized view. The features of the sequences are also studied in order to find a relationship with the best compromise for the bit-allocation between texture and depth data.

1.2 Outline of the thesis

The layout of this thesis divided in four parts. The first part gives the required fundamentals of 3D video. It addresses some basics of human vision, and how the need for depth feeling led to 3D video. The second part addresses the problem of the assessment of virtual rendered views and the specific degradations induced by the synthesis process. The third part proposes two depth map compression approaches, whose main concern is based on the previous observations on the sources of distortions in the rendered views. Finally, the fourth and last part relates the analyses we ran in order to study the relationships between texture and depth data in the context of bit allocation.

Part I - State-of-the-art of 3D Video Communications

Chapter 2 This chapter addresses the origins of the use of illusion of depth together with the fundamentals of human vision. 3D media generation and its display are also discussed. This chapter is motivated by the fact that the knowledge of stereovision and of the many possible display technologies are essential for the understanding of the aim and of the work of this thesis because both elements of the processing chain are dependent on the final quality experienced by the user, thus on the strategic technology choices.

Chapter 3 This chapter is devoted to the issue of 3D video coding. The chapter gives an overview of the coding algorithms for a variety of 3D data representations but it focuses on the compression of MVD data since our research activities concern MVD data.

Chapter 4 In this chapter, the issue of quality assessment when dealing with 3D content is addressed. State-of-the-art subjective and the objective assessment of 3D content quality are discussed.

Part II - Visual quality assessment of synthesized views

Chapter 5 This chapter addresses the principles of view synthesis in the context of MVD and the sources of artifacts related to this process.

Chapter 6 In this chapter, the issue of quality assessment when dealing with 3D content is addressed. In particular, the problem of the evaluation of visual quality of the rendered views is studied through three experiments involving different synthesis algorithms, different objective quality metrics and different subjective evaluation methodologies. The experiments include monoscopic and stereoscopic viewing conditions. This chapter also proposes a hint for the objective quality assessment of synthesized views, based on the results of the experiments.

Part III - LAR-based MVD coding solutions

Chapter 7 This chapter presents the basics of the Locally Adapted Resolution coding method, a still image encoding method whose potential for depth map coding is analyzed in this chapter.

Chapter 8 This chapter presents the novel option we propose for compression of MVD data, based on the studies previously presented on LAR codec performances. The method is meant to preserve edges and to ensure consistent localization of color edges and depth edges. The compression of depth maps is based on the LAR method. The quad-tree representation contributes in the preservation of edges in both the color and depth data. The adopted strategy is meant to be more perceptually driven than state-of-the-art methods.

Chapter 9 This chapter presents a second LAR-based approach for depth maps compression. This method is meant to be more reliable and scalable because it exploits multiresolution repreentation of the depth maps. The prediction technique is based on region segmentation relying on the quad-tree decoded from the LAR stream.

Part IV - Relationships between color and depth data

Chapter 10 This chapter questions bit rate allocation between texture and depth data for

encoding MVD data sequences. The study presented in this chapter includes the compression of MVD data sequences with H.264/MVC in a first case study, and the compression of MVD data with HEVC in a second case study, at different bit-rates in order to determine the best bit rate distribution between depth and texture, according to PSNR measures of the synthesized view. Based on the obtained results, an analysis of different sequence features is proposed to highlight correlations with the best bit-rate allocation.

Part I

State-of-the-art of 3D Video Communications

Every step of the 3D Video processing chain can have an impact on end user visual quality of experience (QoE). Since this thesis aims at providing tools enabling enhanced QoE, the identification of sources of distortions is a primary phase of our research. This step requires the knowledge of the whole 3D Video processing chain. For this reason, this part is devoted to the presentation of 3D Video fundamentals.
In this part, we first address some basics of 3D imaging in Chapter 2. This chapter comes back to the history of the use of illusion of depth and addresses the fundamentals of human vision. A discussion on 3D display and 3D media generation is proposed in this chapter. The next chapter, Chapter 3, concerns an overview of the coding algorithms for different 3D data representations. However, since this thesis focused on Multi-View-plus-Depth data, an emphasis is proposed for the compression of this specific 3D scene representation. Finally, Chapter 4 introduces the basics of our major concern that is the issue of quality assessment of 3D media. Indeed, since our goal is the conception of new tools enhancing the quality of such media, the understanding of the complexity of the assessment task is essential. The chapter discusses the subjective quality assessment methods and the objective quality assessment method, considering the 2D-based approaches and the new trends for 3D media.

CHAPTER 2

3D imaging basics

The final quality experienced by the 3D medium user is dependent on both elements of the processing chain and consequently on the strategic technology choices. This chapter is motivated by the fact that the knowledge of stereovision and of the many possible display technologies are essential for the understanding of the aim and of the work of this thesis. In this chapter, we propose a retrospective glance at the origins of stereoscopy and use of illusion of depth, in a first section. A brief introduction to the fundamentals of human vision is also discussed. The second section focuses on the 3D media generation and its display.

2.1 Principles of Stereoscopic Vision

In this section, we come back to the origins of the use of illusion of depth. Afterwards, we introduce the basics of human vision. Finally, the depth cues that allow human to perceive their three-dimensional environment are discussed.

2.1.1 A brief history of illusion of depth

It is believed that the story of illusion of depth dates back to the Ancient Greece, when Euclid found out that humans perceive the depth thanks to the two different images that we get from the eyes.

At a later time, during the Renaissance, drawing and painting techniques were used to ensure the illusion of depth. Painters such as Leonardo Da Vinci took benefit from the knowledge of human perception to ensure the illusion of depth by using pictorial depth cues, later defined in this chapter.

Later, in 1838, Charles Wheastone [Whe38] patented the Stereoscope device (Fig. 2.1). This system of mirrors allows the simultaneous observation of two slightly different drawings by each eye and the impressive illusion of depth. It is worth noting that this work in stereoscopic visualization was made prior to the invention of photography. With the start of photography, drawings were substituted for photographs.

Afterwards, the first anaglyph images and glasses appeared in the 1850's (the use of two different color filters enables the illusion of depth, see Fig. 2.2). First anaglyph stereo 3D movies appeared in the 1890's. Ever since, the 3D industry has grown up thanks to the advances in 3D graphics for gaming, the advances in display technology, and in 3D content generation.

Nowadays, the term "3D" is often used instead of "stereoscopic image" and this is confusing because the computer graphics community also uses the term "3D" but it then refers to a rendering image of a synthetic scene model. Stereoscopic images are widely studied and used nowadays to provide the illusion of depth, but other technologies such as holography, also known as "true 3D", are still under development.

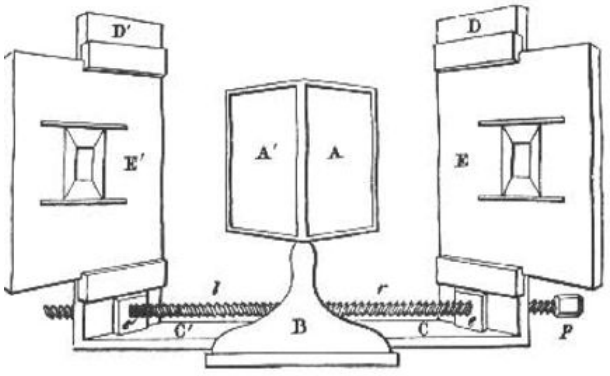


Figure 2.1: *Wheastone stereoscope. Angled mirror A reflects the stereoscopic drawings E toward the viewer's eyes. (Drawing from Bill Gamber and Ken Withers [GW]*

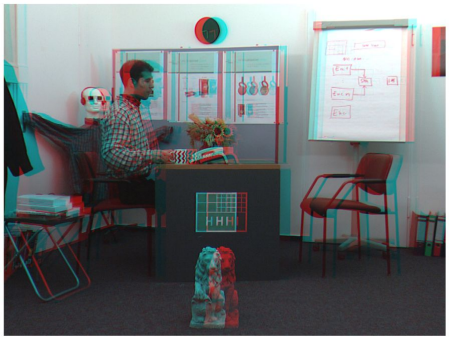


Figure 2.2: *Anaglyph image generated with Book Arrival sequence*

2.1.2 Anatomy

Human vision refers to the very complex neural process that enables the perception of our spatial surroundings. It arouses the interests of researchers because the mechanisms are not clearly understood. The eye is the first organ identified as responsible for vision (Fig. 2.3). The light enters the eye through the cornea, a transparent area, then passes through the aqueous humor to the lens. Then the light goes through the vitreous humor. The retina is sensitive to light. Cells sensitive to strong light, namely cones, are located

at the center of the eye, in a small area called fovea. Cones are responsible for detection of fine details and color vision. Cells sensitive to poor light, namely rods, are located at the outer edges of the retina. Rods are responsible for dark-light vision, and peripheral vision.
At this stage, the light entering cones and rods is converted to an electrical signal, transmitted to the brain through the optical nerves. This signal corresponds to an inverted image of what we see. The visual system is able to correct many errors of the retinal image. The final mental image is sharper thanks to mechanisms defined later in this section, namely vergence and accommodation. Scientists have identified different "visual pathways" that are responsible for different abilities such as the detection of motion, the recognition of objects, etc. For more details on human anatomy, one can refer to [Wan95].

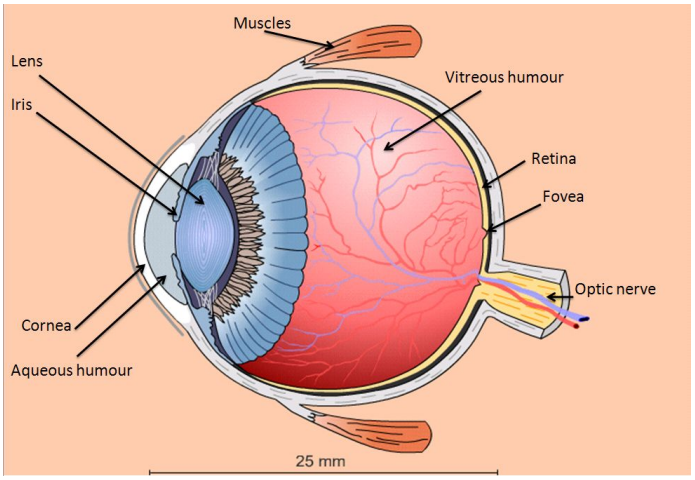


Figure 2.3: *Horizontal section through Human right eye[Per10]*

2.1.3 Human perception of depth

In daily life, we comprehend our three-dimensional surrounding through the use of the complex mechanisms of the human vision system (HVS) [Kau74]. We perceive depth thanks to the two retinal images, but one two-dimensional image already contains a great amount of visual cues for the perception of depth. Artists have understood this for ages, and have taken benefit from these abilities to offer us new representations of the world. The many depth cues are used in various degrees, but stereopsis might be the most impressive and fascinating as it has caught the attention of scientists and of the public since the 1840's [Whe38].

The sources of depth information are generally distinguished in four categories [Pal99]: ocular information, dynamic information, pictorial information and stereoscopic information. These depth cues are discussed in the following.

Ocular information

Ocular cues refer to accommodation and vergence. Both of these phenomena are related to the state of the eye. Accommodation of the eye refers to the act of physiologically adjusting the lens to alter the refractive power and bring objects that are closer to the eye into sharp focus. Vergence is defined as the movement of the two eyes in opposite

direction to enable the fixation to a point or region of interest. These two mechanisms share close links and are interactively solicited [Sch99, PJ90]. Wallach *et al.* [WF71] stated that accommodation and vergence are useful source of depth information to make direct judgment about distance as well as to evaluate the size of objects.

Dynamic information

The dynamic cues refer to depth information from motion. Motion parallax is a dynamic depth cue provided by our motion. As an example, as we move in a car, objects that are close to us seem to go by quicker than objects that are further away. Consequently, motion parallax gives relative information about the distance to an object; it expresses how close an object is from the fixated one. Motion parallax is linked to a visual process, namely the optic flow [Gib50]. It refers to the apparent motion of objects caused by the relative motion between the observer and the scene. Motion can come from moving observed objects or from a moving observer.

Pictorial information

Pictorial depth cues have been applied in visual arts for centuries. They are monocular cues. It means that these cues can be extracted from a sole two-dimensional image (flat image): when closing one eye, you can still perceive the depth of your surroundings. Pictorial depth information includes various cues.

Light and shadow distributions provide knowledge on the shape of the object thanks to the analysis of the reflection of light of its surface (Fig. 2.4(a)).

The interposition or occlusion occurs when a part of the observed object is hidden by another object. In such a situation, the HVS considers the partly hidden object as further away (Fig. 2.4(b)).

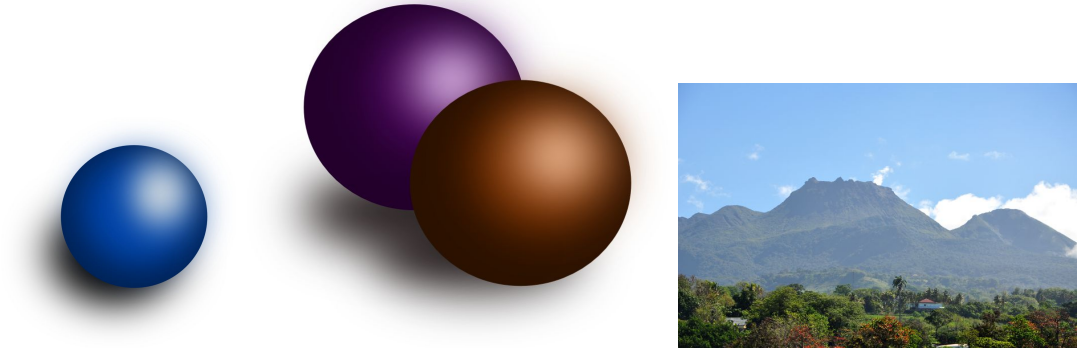
Aerial perspective refers to the fact that the air contains more microscopic particles and moisture, scattering more light when the distance between two object increases (Fig. 2.4(c)).

Relative and known size refers to the fact that a priori knowing the size of objects, objects of similar size seem smaller when placed further away (Fig. 2.4(d)).

Linear perspective is related to vanishing lines or points. It refers to the fact that parallel lines in 3D space converge in the image into a vanishing point (Fig. 2.4(e)). Texture gradient refers to the fact that the size of texture pattern, their density or their orientation provide knowledge on the shape of the object: with distance the texture patterns seem smaller, and their density increases.

Stereoscopic vision

Human eyes are separated by 6.3 cm on average. Due to this lateral displacement, we get two slightly different images of the same scene (disparate images) from each eye. This is known as binocular disparity. The brain processes these two images to render a depth appreciation. This phenomenon is a binocular depth cue and is referred to as stereopsis.



(a) Light and shadow (b) Occlusion (c) Aerial perspective

(d) Relative size (e) Vanishing lines

Figure 2.4: *Depth information from monocular cues.*

The word stereopsis comes from the Greek *"stereos"*, solid or firm, and *"oyis"*, look or appearance.

Fusion refers to the neural process that forms one single image out of the retinal images coming from the two eyes. If there are matching features in both images, fusion is possible. Otherwise, phenomena such as binocular rivalry, suppression or superimposition may occur. Binocular rivalry [Bla89] refers to the alternating perception of the two images. Suppression refers to the elimination of one image. Superimposition refers to the fact that one image overlaps the other.

The distance, in horizontal direction, between the corresponding points of the two images is referred to as retinal disparity. If the eyes converge on an object, the resulting corresponding points of the retinal images will have "zero disparity". In this case, these points lie in an area called horopter (see Fig. 2.5). Points lying in an area close to the horopter called Panum's fusional area, are fused perceptually into a single experienced image. In that case we see two slightly different images but you do not experience double vision as illustrated on Fig. 2.5. The interpretation of retinal images to produce stereopsis is entirely mental, and must be learned: the stereoscopic ability is not present at birth.

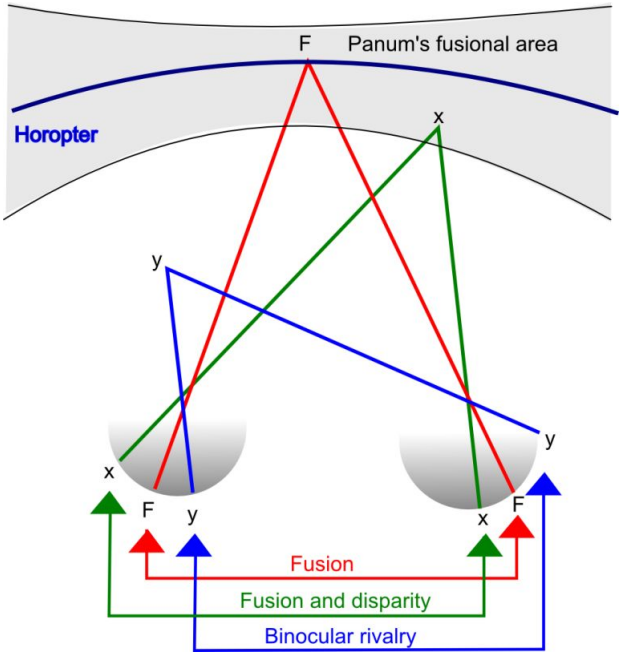


Figure 2.5: *Basics of stereoscopic viewing [Pat07]*

Ocular cues, dynamic cues and pictorial cues can be defined as monocular cues. Stereoscopic cues are defined as binocular cues because they require the two eyes. Depth perception is the result of the use of many cues with various degrees, as shown by Cutting and Vishton [CV95] and on Fig. 2.6. Besides, in this study ([CV95]), the authors showed that occlusion is dominant over all other cues, and is only approached by binocular disparity, as well as the small effect of accommodation and convergence. Although the complex mechanisms of human vision are not clearly understood, the use of monocular and binocular cues already enable creators and artists to impress the public by illusion of depth. The next section provides an illustration of this use of illusion of depth, in the case of 3D Video, through 3D contents and displays.

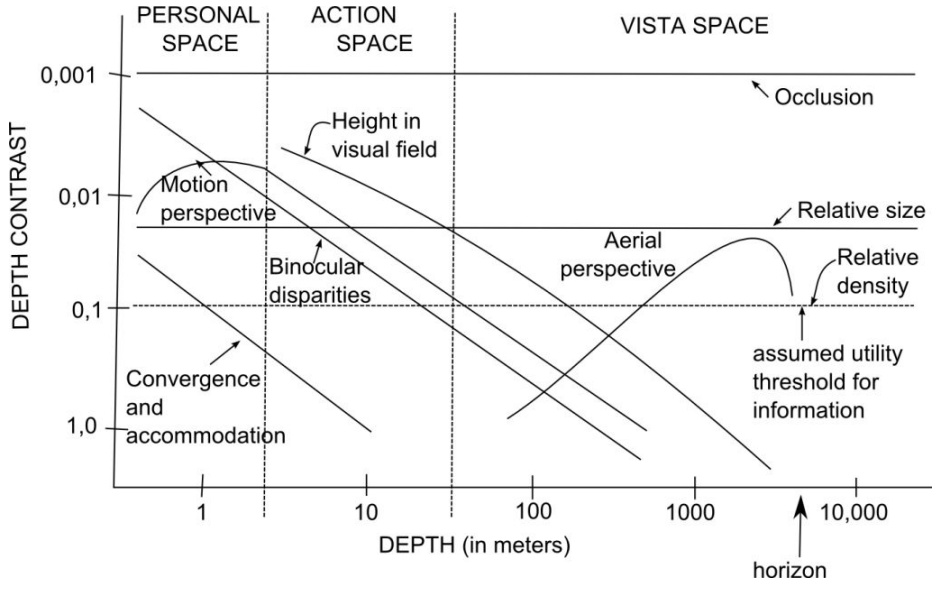


Figure 2.6: *Cues effects on depth perception [CV95].*

2.2 3D content generation and display

Many solutions have been found in order to emulate depth through media. The expected depth effect is obtained thanks to prepared 3D contents and adapted displays. In this section, the generation of 3D video sequences is first addressed. Then, the display technologies and their limitations are introduced before presenting the different possible 3D scene representations from the acquired videos.

2.2.1 Content generation

The stereoscopic visualization of a 3D media requires the use of, at least, two slightly different images, namely a stereopair. This stereopair is either originally acquired by two different acquisition devices or it is artificially generated from a previously acquired viewpoint. The generation of a new color image is possible thanks to the use of extrapolation or interpolation algorithms, from color data, or thanks to Depth-Image-Based-Rendering (DIBR) algorithms [Feh04]. This newly generated color view is also referred to as the virtual view, or the synthesized view. DIBR algorithms require a depth map of the scene. A depth map is a monochromatic image whose pixels indicate the distance of the corresponding color pixel to the acquiring camera.

Most of the 3D sequences are shot through multiple cameras, two cameras being a minimum. However, due to recent needs, 2D conventional sequences can be converted to multiple view sequences thanks to the development of 2D-3D conversion algorithms [OMT+96]. These algorithms usually include a segmentation step in order to generate a depth map from the 2D color image [YYED11].

When the 3D media is acquired by multiple cameras, two configurations of acquisition are possible [Yam06]: the parallel camera configuration and the convergent camera configuration, also called toed-in camera configuration (Fig. 2.7). In the parallel configuration, the zero-disparity point is at infinity. Objects near the camera cause visual discomfort because their binocular disparities can be large. In a convergent configuration, the zero-disparity point is at a finite distance. Although, the absolute disparities can be smaller in this configuration, it may lead to distortions such as keystone and vertical disparities [WDK93]. The choices result in a trade-off taking into account all possible distortions [Yam97]. In [CFBLC11], the authors recommend new shooting rules considering both stereoscopic distortion and comfortable viewing zone. They state that the most important point is to guarantee the perceived scene range is within the comfortable viewing zone by adapting the scene parameters or camera parameters.

As said above, a typical depth map is a monochromatic, luminance-only video signal. It is a grey scale image with smooth areas and sharp edges. The depth range goes from Z_{near} to Z_{far} and is quantized with 8 bits. The lighter the pixel (value 255), the closer to the optical center of the camera. Depth maps can be acquired through several methods either from monocular videos, either from multi-view videos or from specific cameras. One method relies on the measure of the real 3D properties of the scene objects by using a range finder such as LIDAR system (Light Detection and Ranging) [Sch10]. Depth maps can also be estimated by using stereo-correspondence-based algorithms when two adjacent color views are available ([SS02, KAF+07, WZ08, YWY+06]). Another way for generating depth maps consists in extracting the motion vector information, or the optical flow [Sou10]. Other methods are based on image classification and vanishing lines or feature

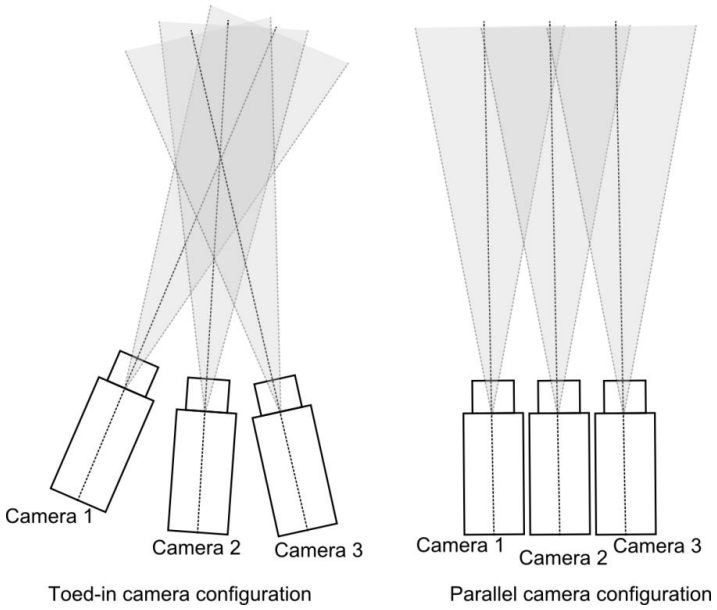


Figure 2.7: *Camera configurations for 3D media shooting. Toed-in camera configuration (left) and parallel camera configuration (right).*

points detection [BCCLC04]. Specific cameras, called "Z-cameras", are able to capture the per pixel depth with the video, but the quality of the obtained depth is limited. Depth data is usually useful in 3D visualization because it allows manipulations on the parallax baseline. The parallax baseline refers to the shift or the movement directly dependent on the distance between two adjacent viewing cameras. However, although powerful solutions are available for depth estimation, estimation errors remain. Such errors induce artifacts in the virtual synthesized view. A smooth, accurate and reliable depth system is still under investigation and the impact of this source of distortions has to be studied.

2.2.2 3D imaging displays

Among the possible so-called 3D applications, the following may be targeted: 3D-TV for home entertainment, providing the user with a feeling of immersion; video games; training for junior professionals ([IPLW07] in the medical field, for instance); free-viewpoint video (FVV) which allows the user to freely navigate inside the scene by selecting the viewpoint of the video scene; special video effects (such as the effects used in "The Matrix" when freezing time while moving around objects). These applications are achievable thanks to the display technology and thanks to an adequate choice for the data representation. Many representations exist, both have their advantages and drawbacks that are discussed in this and the following section. As mentioned in 2.1.1, the concept 3D display has a long history and there exists many different methods to allow 3D viewing [BWS+07, RHFL10]. The technologies have been divided as follows: multi-view displays with fixed viewing zone, multi-view displays, integral imaging displays, volumetric displays and holographic displays.

Binocular displays with fixed viewing zone

Multi-view displays with fixed viewing zone are so called because they provide a fixed viewing zone for each eye. Consequently this device allows only one user to experience

depth. It is the simplest type of display since a single stereopair is provided. Different types of multiplexing methods can be used to produce the viewing zone of each eye: lenticular type, parallax-barrier type, wavelength-division (anaglyph type), time-division, polarization-division type or combinations of it. Glasses may be needed to experience the illusion depth (passive glasses: anaglyph or polarization; active glasses: LCD-shutter glasses). In that case, the device is described as stereoscopic. Otherwise, when there is no need for glasses, the device is described as autostereoscopic.

Multi-view displays

Multi-view displays provide viewing zones for several users contrary to multi-image displays with fixed viewing zone. The device provides a set of perspective views in the viewing field, and allows a certain range of motion parallax when the users move to adjacent viewing zones. However, the depth range available in such devices is limited [CFBLC10], and the image resolution is reduced according to the number of views rendered by the device. Multi-view displays are already marketed despite their limitations in the quality of the stereo effect.

Volumetric displays

Volumetric displays create an image in which each point of a scene reaches its actual position in space. The scene is reproduced within a volume of space that allows a wide range of viewing angles for the observers. This technology calls a more natural viewing than binocular displays because the eye can focus at a real point. Volumetric displays can employ focused or intersecting laser beams to create voxels, or laser beams or layered images on moving screens. Although these devices provide a wide range of viewing their main drawback comes from their image transparency. This technology is not available for mass market yet but several companies are involved in its improvement.

Integral imaging

Integral imaging displays are autostereoscopic. They use an array of small lenses that are either spherical or cylindrical in front of an image. This produces a light field that makes a lens looking different depending on the viewing angle. Not only stereopairs are available, but it allows motion parallax as the observer moves. This technology is expected to be the subject for mass commercialization in the future years.

Holographic displays

Holographic displays rely on diffraction-based coherent imaging methods. They can reproduce the wave field of a 3D scene in space by modulating coherent light. Holographic display is often called as “true 3D”. However, due to the numerous issues to overcome, this technology is not mature yet, and is not ready for mass market.

2.2.3 Limitations of the 3D display

A 3D media can be experienced through various types of displays as discussed above. Over the past decades, research efforts focused on the development of stereoscopic imaging systems but some fundamental issues remain unsolved, like the conflict of accommodation and vergence.
As defined in 2.1.3, vergence refers to the convergence of the fixation axes of the two eyes.

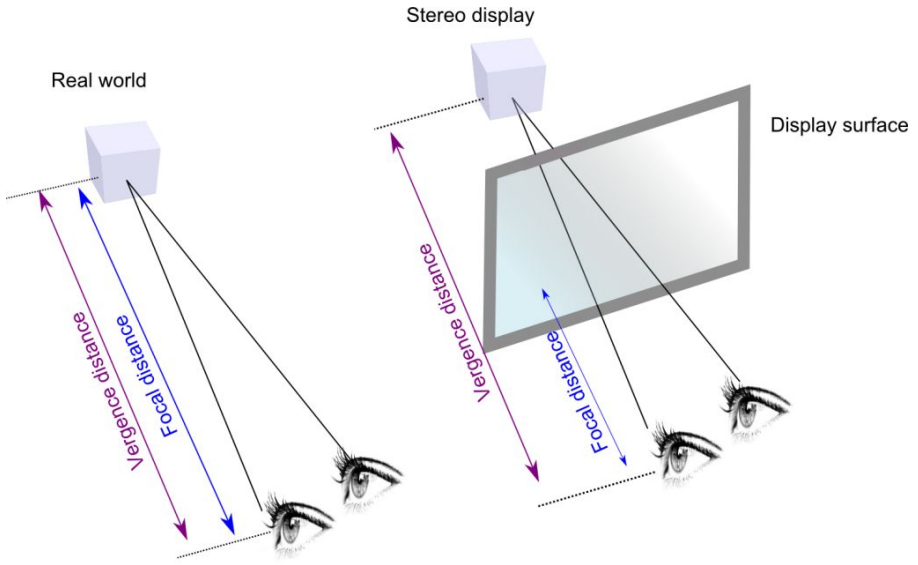


Figure 2.8: *Natural viewing (left) and stereoscopic viewing with a 3D stereo display (right). Normal viewing and also holographic displays provide correct vergence and accommodation stimuli when fixing on an object. A stereogram fails in providing correct vergence and accommodation stimuli unless the fixed object is at the focal distance, on the display surface.*

Focal distance is the distance to which the eyes are focused, this refers to the accommodation. In natural conditions, vergence and focal distance are the same. When viewing an object through a stereoscopic display, the display surface is nearer than the object, so focal and vergence distance mismatch, as illustrated in Fig. 2.8. Studies [BAHG08] highlighted that this phenomenon causes visual discomfort and eyestrain.

In addition to the vergence-accommodation conflict, there are other unsolved issues [LIH07, WDK93]. Crosstalk can occur when the images are not perfectly separated. This term designate the ghosting effect experienced when left and right views are not correctly separated and seem superimposed. Its effect varies depending on the position of the observer and on the quality of the optical filter.
The depth rendering ability is also another limitation of the 3D displays. In [CFBLC10], Chen *et al.* point out the issue: perception of stereoscopic depth is dependent on the 3D content, on the viewing distance and on the display characteristics. In [CFBLC10], Chen *et al.* propose an analysis of depth rendering abilities of different displays according to their characteristics. In [Pat07], recommendations for 3D display design are stated and the authors include, among others, concerns regarding the interocular crosstalk (to limit the chromatic aberration), the spatio-temporal properties of the display (to enable a sufficient depth range), the frame rate (to avoid the perception of flicker), the viewing distance (to enable the depth perception).

2.2.4 3D data representation

Various 3D scene representation formats exist in different 3D video systems and applications. These formats involve various types of data, such as multiview video, and geometry data in the form of depth or 3D meshes. The main requirements for a typical 3D scene

representation are its ability to provide an exhaustive and reliable description of the 3D scene, in preparation for the rendering of any viewpoint within this scene; its resilience to compression degradations; its low storage capacity or its easiness of compression. In this document, only image-based formats are discussed, i.e., only methods that rely on a set of two-dimensional images representing the 3D scene. Fig. 2.9 illustrates the different formats of 3D data.

Stereoscopic video

The stereoscopic video consists in a pair of 2D conventional video sequences, one acquired for the left eye and the other one acquired for the right eye. Although it might be the simplest type of acquisition method for a 3D media, it contains many drawbacks: as the baseline depends on the acquisition configuration, the depth effect can hardly be modified. Even by estimating a disparity map from the two views, the parallax is limited. Thus, applications such as FVV are not feasible in this case. As for the 3D display, the most appropriate is the multi-view displays with fixed viewing zone, in this case.

Video-plus-depth data

Video-plus-depth, also denoted as 2D+Z, data representation consists in a conventional 2D video and its associated depth video.The conventional 2D video is generally referred to as color or texture video. It is an alternative to stereoscopic video because an artificial stereopair can be rendered by DIBR techniques [Feh04] from the video and depth information. However, since only one color view is available, occlusions can be hardly handled by the synthesis process when the baseline increases: when the distance between two viewpoints increases, objects that were not visible in the base viewpoint become visible in the novel viewpoint. Since only the color information from the base viewpoint is available, discovered areas to be fulfilled are prone to errors. So, the baseline range is limited with such data representation and applications such as FVV is not feasible.

Multi-view video data

Multi-view video (MVV) is considered as an extension of stereoscopic video since the number of conventional video sequences is generally higher than two in this case. Since multiple views are available, special devices can display multiple views simultaneously. This allows head motion parallax viewing. However, the amount of data to be processed increases compared to conventional stereo video sequences. Moreover, if the generation of a novel viewpoint is required, the quality of the synthesis is limited since only the color information is available. However, depth maps can be generated from MVV data. As for the display, multi-view displays are appropriate, providing the images are adapted or adjusted to the display characteristics. Application such as integral imaging or autostereoscopic experience for multiple users are possible with this 3D scene representation.

Multi-view video-plus-depth data

Multi-view video-plus-depth (MVD) data is considered as an extension of the video-plus-depth data representation because it consists in a set of conventional 2D video sequences and a set of corresponding depth sequences. It was created to overcome the limitations of the previous 3D data representations, when MPEG started an activity related to this issue [SMM+06]. Since depth information is available for multiple viewpoints, the distance range from the virtual viewpoints to base viewpoints is larger. This 3D scene representation

allows FVV applications. However, as it will be discussed later in this document, both the amount of data to be processed and the complexity increase.

Layered depth video data

Layered depth video data (LDV) is an alternative to MVD data representation. The layered depth image (LDI) [SGHS98] data representation is the base component which LDV is temporally extended from. Instead of storing an image, thus a 2D array of depth information, the LDI stores a 2D array of set of pixels along viewing lines, sorted from foreground to background. So the front samples represent the first surface seen along that viewing line; the next pixels represent the next surface seen along that viewing line, etc. Compared to MVD, LDV is meant to reduce the inter-view redundancies and carry a lower amount of data. However, the quality of the synthesized view in the case of use of LDV, is widely dependent on the space sampling used for creating the LDV. As for the display and applications, LDV can target the same as MVD, since depth data is available at various viewpoints.

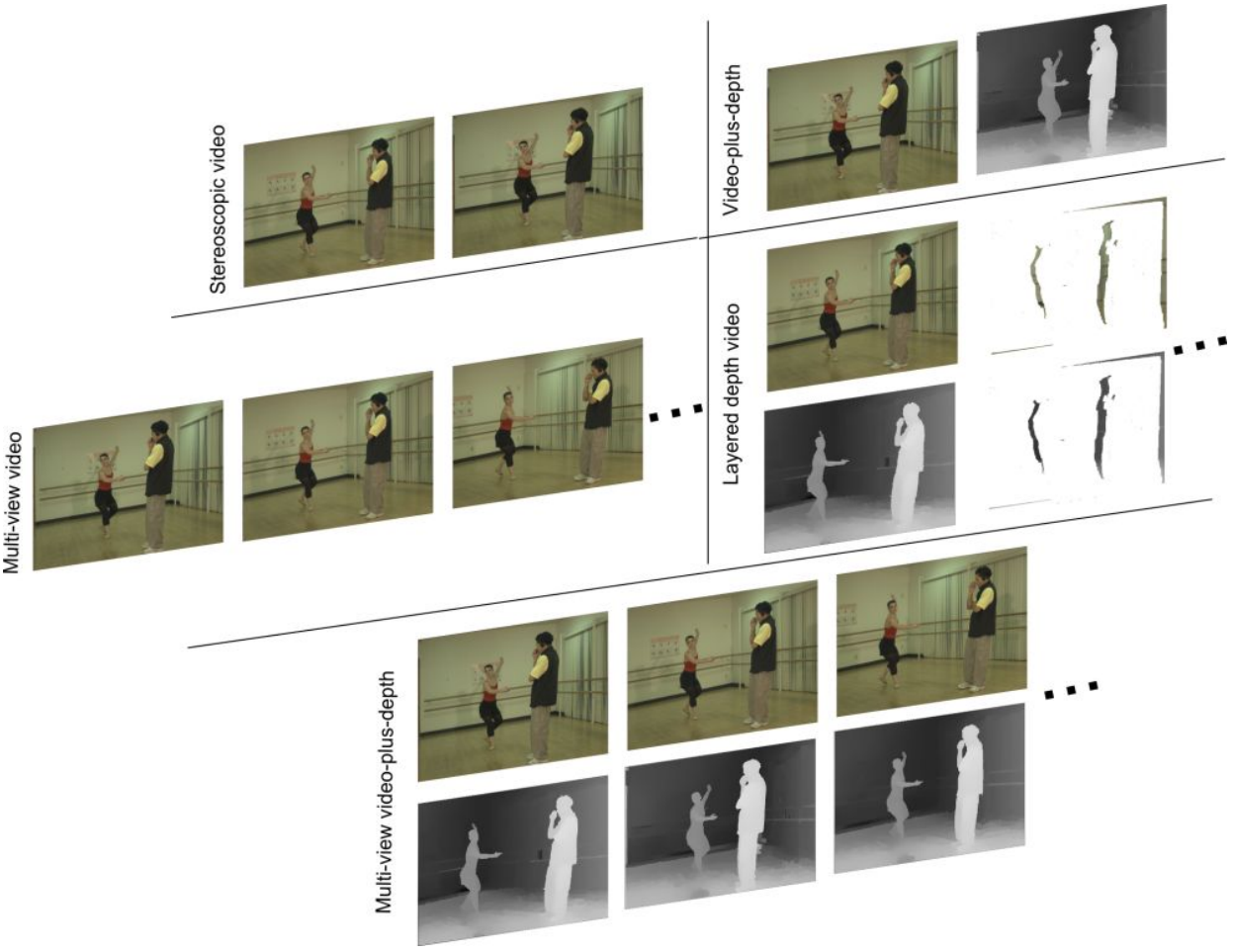


Figure 2.9: *3D data representations.*

2.3 Conclusion

In this chapter we presented some basics of 3D imaging. The understanding of stereoscopic vision is essential both for the design of subjective quality assessment tests and for the conception of coding frameworks. The use of tools to mimic depth feeling has been known for years. Nowadays, with the improvements of hardware, emerging 3D display solutions have arisen. They still suffer some limitations, that shall be overcome for the sake of 3D Video success. In addition, 3D contents have to be created and store. This brings the issue of 3D scene representations. Various representations have been considered in this chapter. The choice for the appropriate 3D scene representation widely depends on the expected case of use. Considering the advantages and the drawbacks of the presented 3D scene representations, we chose to focus on MVD data. As explained, in that case, the amount of data to process is significant and the need for adapted coding method must be addressed. The next chapter presents basics of 3D video coding in the following, with an emphasis on MVD coding because our contributions are in line with this field.

CHAPTER 3

3D video coding

This chapter is devoted to the issue of 3D video coding. In the first part, an overview of the coding algorithms for the aforementioned 3D data representations is presented, as an introduction. The rest of the chapter focuses on the compression of MVD data. Since there is no standardized framework up to now for this 3D data representation, this chapter reviews the most commonly used methods in the two following parts. Afterwards, the methods proposed in the 3DV Group of MPEG call for proposals are addressed.

3.1 Overview of coding algorithms for 3D contents

The previous chapter presented different types of 3D data representations, with a focus on image-based representations. 3DTV systems are numerous and can rely on various types of data. This huge amount of data needs to be compressed and most of the proposed coding methods rely on the extension of available classical video coding frameworks. For most of the presented 3D data representations, encoding methods have already been standardized by the Moving Picture Experts Group (MPEG). MPEG is an international standardization committee. Its role involves the assessment of proposals with a view to adopt standardized frameworks in the field of digital media. MPEG was formed by the ISO (International Organization for Standardization) in 1988.
The main concern of the committee is to ensure backward compatibility with existing systems. This section reviews the available standards for the different 3D representations.

Stereoscopic video

Because of their redundancies, it seems natural to predict one of the two views through the other. Numerous standards have been developed for stereoscopic video. The most commonly used is known as the Multiview Profile (MVP)[BT.98], defined in ITU-T Rec H.262/ISO/IEC 13818-2 MPEG-2 [HPN97]. The efficiency of the method relies on the exploitation of inter-view and temporal redundancies thanks to prediction tools. The left eye view is encoded with temporal prediction only, using standard MPEG-2. The right view is encoded through temporal prediction and also from the left view available data. So, inter-view prediction is used relying on the base layer (the left view) to encode the second layer (the right view), as illustrated in Figure 3.1. This fulfills the backward compatibility

constraint with the Main Profile of H.262/MPEG-2 Video because it is possible to decode only the left view from the bit stream and display a conventional 2D video.

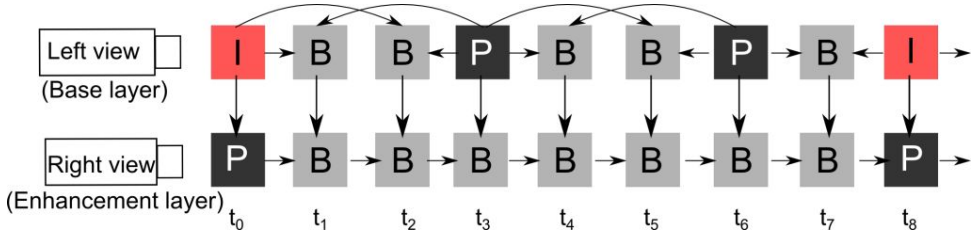


Figure 3.1: *Prediction structure in H.262/MPEG-2 MVP.*

Inter-view prediction is based on the same principles as motion estimation and compensation in MPEG-2. This is justified by the fact that the disparity between each view is considered as equivalent to a dense motion field in between two consecutive images of a video sequence. However, it should be noted that there are differences between motion compensation and disparity compensation. First, the disparity vector fields are different from the motion vector fields, because disparities are relatively large, depending on how close a 3D point is to the camera. Second, views of the stereopair are generally less similar than temporally adjacent frames of a video sequence, because of the importance of the discovered areas. In other words, the higher the baseline distance between two views the lower the gain from inter-view prediction.

Video-plus-depth data

In the ATTEST project[FKDB+02], a method ensuring backward compatibility with respect to DVB enables the compression of 2D+Z data representation. It is based on the assumption that only 10%–20% of the bit rate which is necessary to encode the texture video is sufficient to encode the depth at good quality. The texture video sequence is the base layer, encoded with standard MPEG-2 to ensure the backward compatibility. The additional layer contains the encoded depth data. Then, MPEG specified a similar format "ISO/IEC 23002-3 Representation of Auxiliary Video and Supplemental Information," known as MPEG-C Part 3[JTC07], for 2D+Z data representation. Compared to stereoscopic video data, the total bandwidth for 2D+Z data transmission is reduced.

Multi-view video data

As presented in section 2.2.4, MVV data refers to multiple views of the same scene. The Multi-view Video Coding (MVC) standard already addresses the compression of such data representation. Exhaustive experiments were conducted within ISO MPEG standardization working groups to define the framework having the best performances [MSMW07c, FMG07, MMSW06]. MVC exploits combined temporal/interview prediction. Prediction of images is performed from temporal neighbor images and from adjacent neighbor images at the same instant of time. MVC is based on a state-of-the-art video codec H.264/AVC[WSBL03] that supports hierarchical B-prediction, as illustrated in Figure 3.2. In this figure, "camera 0" is the base view. In [SMS+07], the authors state that although the exploitation of temporal/interview redundancies outperforms the simple independent encoding of multiple streams, the gain is dependent on the content and on the acquisition configuration (baseline distance of acquiring cameras, motion, texture, etc).

In this paper, the authors state that the gain in terms of peak-signal-to-noise ratio was 0.5dB and below.
For the interview prediction, MVC supports the definition of views dependencies through the sequence parameter sets (SPS) syntax.
To ensure backward compatibility, a *base view* is defined and is independently coded. Thus, it can be decoded without the use of the adjacent neighbor views, since only temporal prediction is used in that case.
Note that MVC can also be used to encode the Stereoscopic Video sequences since stereoscopic video is a special case of MVV with two views only.

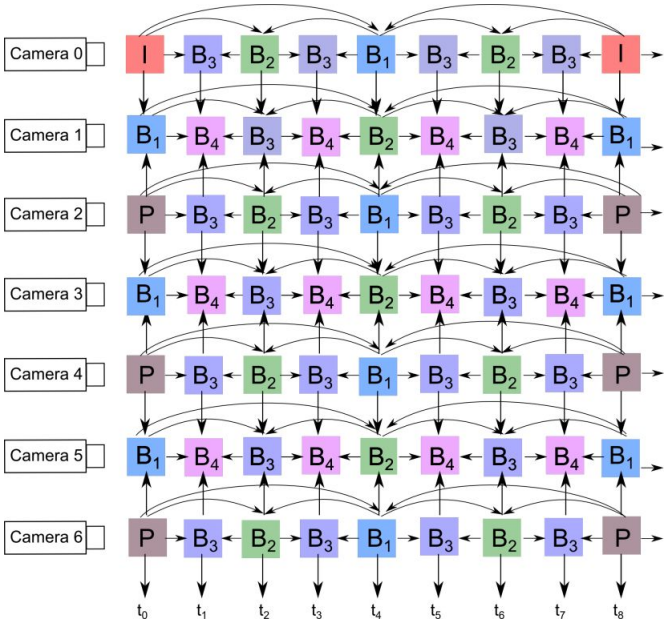


Figure 3.2: *Prediction structure in MVC, using temporal and inter-view predictions.*

Emerging standards for other 3D data representations

Other compression frameworks dedicated to 3D data representations are still under investigation. As for an example, there is no standardized coding method for LDV, presented in section 2.2.4, the alternative representation of MVD. Most of the proposed approaches for this data representation, are based on MVC. Compared to MVD, the amount of data to be encoded is lower because of the properties of LDV. The difficulty consists in encoding the holed images, typically, the back layers of LDV. In [YH07], the color, the depth and the image of a number of layers per pixels are encoded. The authors tested two alternatives: the first one consists in filling the holes with the pixels of the first layer and then eliminate them automatically, knowing the number of layers per pixel; the second one consists in aggregating the pixels horizontally and then aggregate both layers. LDV representation has not reached yet the same interest as MVD has, considering the expended energy of standardization activities dedicated to the latter. LDV has actually been considered by MPEG, but due to the results of the core experiments, the efforts have been oriented to MVD. Encoding methods are still to be proposed.

3.2 MVD coding

This section is dedicated to MVD coding. This emphasis is justified by the fact that this thesis focused on MVD coding.
Currently, a compression framework for MVD is under standardization within ISO MPEG. This section will review pioneering studies in this area, in a first part. Then, a second part will present the tools that are currently evaluated in the MPEG standardization context.

3.2.1 Pioneering studies

A first attempt of coding design is based on the assumption that depth maps can be considered as conventional monochromatic images. This assumption allows the use of MVC standard for both texture and depth data, separately[MSMW07b]. However, since MVC is the video coding standard for MVV data, it does not originally involve the transmission of depth sequences. Thus, redundancies between texture and depth data, as well as interactions between texture and depth data artifacts are not taken into consideration. When using MVC for encoding texture and depth data, separately, Merkle *et al.* observe in [MSMW07a] that depth data coding quality has a strong influence on the quality of the rendered intermediate view. In particular, distortions occur around depth discontinuities, located at borders of objects. In this study, the ratio between rates for depth and texture is kept constant: 75% of the total rate is dedicated to texture and 25% is dedicated to depth data. It is worth noting that this choice is questionable since the influence of depth quality on the visual quality of rendered views was not previously studied. This choice is motivated by the assumption that being monochromatic, depth requires a budget about three times lower than that of the texture.

The observation of the quality of the rendered views from compressed texture and depth data[MFdW07], rather than the quality of texture only or depth only has led to the consideration of the relationships between texture and depth in preparation for the synthesis process. It appears that depth maps need to be considered as non-natural images since depth values represent geometrical 3D positions of scene points. Extending the application of 2D conventional codecs on depth data leads to artifacts that may be imperceptible when visualizing the depth map, but they produce distortions on the synthesized view. Indeed, when performing a synthesis, the warping process relies on wrong depth values, because of the unadapted quantization. The impact of depth compression on visual quality of synthesized views can be explained by the fact that 2D codecs are optimized for human visual perception of color images. Consequently, the pioneering studies on MVD compression led to the observation that depth data require a specific compression method. This method should ideally 1) exploit the redundancies between texture and depth, and 2) be optimized for the enhancement of the visual quality of the synthesized view.

Efforts have been directed in order to propose depth compression methods more adapted to the special features of depth maps. Depth maps contain smooth areas and sharp edges. This observation has motivated the choice for content-adapted methods. Morvan *et al.*[MdWF06] have proposed to represent the depth map thanks to platelets (piecewise linear functions). The depth map is first divided through quad-tree decomposition and each block is approximated by a platelet. The platelet-based compression outperforms JPEG2000 in the study. However, in this study, the gain is evaluated with respect to the depth distortion (in PSNR). This protocol of validation is questionable because since the artifacts in the two compared methods are different, their impact on the synthesis

may also be different. Yet, the quality of the synthesized views generated from the decoded depth maps is not presented. Graziosi *et al.*[GRP+10] have also proposed a block partitioning method associated to least-square prediction for depth map compression. In this method, the validation is also achieved by comparing the depth map distortion for different compression scheme (JPEG2000 and H.264 intra). The method includes the use of a dictionary, containing concatenations of scaled versions of previously encoded image blocks.

The exploitation of redundancies between texture and depth maps has also been the object of various research activities. In [DTP09], Daribo *et al.* questioned the principle of fixed bit-rate distribution between texture and depth. The authors have proposed an H.264-based algorithm that uses a joint estimation of the motion vector field for texture motion information and for the depth map sequence. The proposed bit-rate allocation method is based on a rate-distortion criterion, relying on the distortion of depth and on the distortion of texture, but not on the distortion of the resulting synthesized view. Though, as mentioned before, since depth maps are not natural images, the independent analysis of depth maps distortion and of texture distortion may not be sufficient.

For that matter, encoding algorithms supporting rate allocation between depth and texture have been studied, with a focus on the optimization of the synthesized view quality. In [MFdW07], Morvan *et al.* proposed a bit-rate allocation method based on the analysis of the synthesized view quality. However, the quality criterion is the Mean Squared Error (MSE), which is known for its limited ability to express visual quality. In that study, the coding method is H.264/AVC and the proposed algorithm selects the quantization parameter for texture and that for the depth map according to the MSE score of the resulting synthesized view.

3.2.2 Tools under standardization

ISO/IEC JTC1/SC29/WG11 (MPEG) issued a Call for Proposals on 3D Video Coding in March 2011. Considering the influence of depth estimation and view synthesis on the quality of the reconstructed views, MPEG called for contributions in such areas in addition to the definition of a 3D data format and its compression method. The Call for Proposals includes two different tests categories:

- AVC compatible: the proposed method should include forward compatibility with AVC.
- HEVC compatible: the proposed method should include forward compatibility with HEVC, or no constraints.

Among the 22 proposals (submitted in November 2011), only one used LDV and only one used MVV. All other proposals were designed for MVD data.
This section will focus on tools meeting the requirements mentioned above regarding the exploitation of texture/depth redundancies, and the optimization of visual quality of the synthesized view.
However, some terms need to be defined first. An access unit is defined as the group of images consisting of texture and depth view components at time t, as illustrated in Figure 3.3. To enable AVC-compatibility, or HEVC-compatibility (depending on the test category), one texture view component is defined as the AVC-compatible *base view* (re-

spectively HEVC-compatible *base view*). Other views are called *enhanced views*. The following tools are mentioned in [1112].

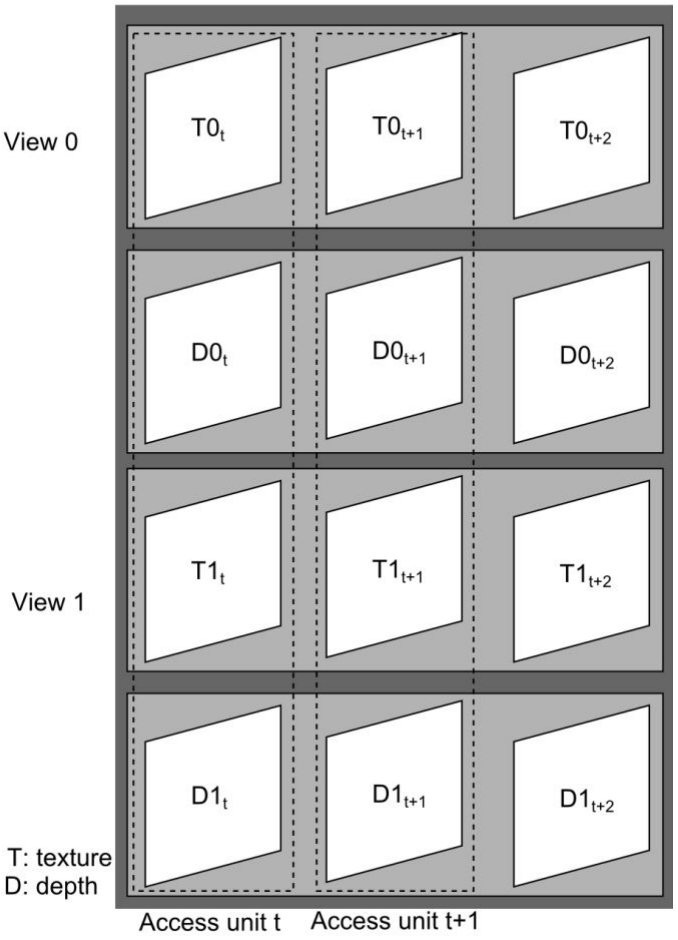


Figure 3.3: *Definition of an Access Unit.*

View Synthesis Prediction (VSP)

View Synthesis Prediction (VSP) is used for coding both texture and depth map data. The current view (an enhanced view) can be encoded using view synthesis prediction from previously coded texture and depth view components of the same access unit. A predicted picture is synthesized by warping the image signals of the reference pictures into the target viewpoint, corresponding to the predicted picture viewpoint. While block-wise disparity compensated prediction was already present in MVC with a limited accuracy, VSP provides improved predicted frames. This can be explained by the fact that VSP takes scene geometry into consideration, by using DIBR algorithms. Details on the principles of VSP can be found in [YV09]

Adaptive Depth Quantization

Adaptive Depth Quantization is proposed to enable the adjustment of depth quantization according to the corresponding texture information. The choice for the adaptive depth quantization parameter is computed through an objective quality metric and the analysis of the texture complexity of the blocks. The objective metric relies on the sum of squared distance of the reconstructed depth map and on the analysis of the view warping through the knowledge of the reconstructed texture, the reconstructed depth map and Z_{near} and Z_{far} values. The smoother the texture, the coarser the quantization step size.

In-loop joint inter-view depth filtering (JVDF)

JVDF allows depth filtering by a weighted average of two depth maps: the first depth map contains the original depth value at a given viewpoint, say viewpoint 2; the second depth map is the result of the projection of the depth map from another viewpoint, say viewpoint 1, to the target viewpoint, that is to say viewpoint 2. This is expressed as follows:

$$\hat{z_2} = w_1.z_{1\to 2} + w_2.z_2 \tag{3.1}$$

where, $\hat{z_2}$ is the filtered depth map, z_2 is the original depth map, $z_{1\to 2}$ is the depth map warped from viewpoint 1 to viewpoint 2, and w_1 and w_2 are the weighting coefficients. The latter are described by:

$$\begin{array}{l} \text{If} |z_{1\to 2} - z_2| < Th, w_1 = w_2 = 0.5 \\ \text{Otherwise}, w_1 = 0, w_2 = 1 \end{array} \tag{3.2}$$

where Th is a threshold value that has to be transmitted to the decoder.

Depth-based motion vectors

Depth-based motion vectors use available depth map data and utilize it for coding and decoding of associated texture data. This coding tool is enabled for enhanced texture coding and requires depth map data to be coded prior to the texture data.

Reduced resolution coding

Proposed tools for MVD data compression consider the coding of the original data (both texture and depth) at different spatial resolutions. As for an example, one of the proposals suggests the following. It considers three different possible spatial resolutions: full resolution, half resolution and three quarter resolution in horizontal and vertical directions. For a given target bit-rate, the first Access Unit of the input test sequence is coded for each of the three possible resolutions. The resolution giving the lowest mean squared error (MSE) is selected and the input test sequence is rescaled prior to coding. Rescaling is performed at the decoder side, thanks to the provided dimension information from the encoder.

Compression of camera parameters

Since view synthesis prediction is planned to be included in the MVD coding framework, camera parameters and depth range ($[Z_{near}$, $Z_{far}]$ have to be encoded. Each view has its own camera parameters and its own depth range. When cameras move along the sequence, camera parameters and depth range change. Proposals cover this issue by using flags to indicate the updated components, and encoding only the latter.

3.3 Conclusion

Research on coding of 3D contents (stereoscopic video, multi-view video, etc.) has led to international standards, some of which have been reviewed in this chapter. Despite the efforts and the good level already reached in 3D coding, improvements are still under study. Concerning the MVD data representation, a compression framework is still under normalization, and should be released by early 2013. In line with this research thematic, the issue of depth maps compression will also be addressed in the remaining of this thesis. Since any method needs to be validated by assessment tools, the following chapter introduces principles of quality evaluation of 3D video sequences.

CHAPTER 4

Quality assessment of 3D video sequences

This chapter aims at highlighting the issue of quality assessment when dealing with 3D content. As an introduction, the first section gives an overview of the most common issues. The second section addresses subjective assessment of 3D content quality. Finally, the third section discusses objective assessment of 3D content quality.

4.1 The peculiar task of assessing 3D contents

3D video applications have encouraged numerous investigations for various applications (see 2.2.4). The most popular applications can be considered as 3D-TV and FVV. 3DTV provides a depth feeling thanks to an appropriate 3D display. FVV interactively allows the user to control the viewpoint in the scene. Considering the demand for high-quality visual content, the success of 3D video applications is closely related to its ability to provide viewers with a high quality level of visual experience. While many efforts have been dedicated to visual quality assessment in the last twenty years, some issues still remain unsolved in the case of 3D video. The assessment of 3D contents arises different issues:

- **Evaluation of the synthesized views.** 3D-TV as well as FVV require view synthesis. This process is known as DIBR and can induce new types of artifacts, that will be discussed later in this document. Since view synthesis is fundamental for both 3D-TV and FVV, the evaluation of the synthesized views quality is crucial.
- **Specific distortions in DIBR.** Artifacts in DIBR are mainly geometric distortions. These distortions are different from those commonly encountered in video compression, and that are assessed by usual evaluation methods: most video coding standards rely on DCT (Discrete Cosine Transform [ANR74]), and the resulting artifacts are specific (some of them are described in [YW98]). These artifacts are often scattered over the whole image, whereas DIBR related artifacts are mostly located around the disoccluded regions. Thus, since most of the usual objective quality metrics were initially created to address usual specific distortions, they may be unsuitable to the problem of DIBR evaluation.
- **Case of use and visualization.** The evaluation of DIBR systems is a difficult task because the type of evaluation differs depending on the context of use. It is not

the same factors that are involved in all of the 3D imaging applications. A major discriminatory factor is the stereopsis phenomenon (fusion of left and right views in the human visual system, as defined in 2.1.3). This is used by 3D-TV and this reproduces stereoscopic vision. This includes psycho-physiological mechanisms which are not completely understood. A FVV application is not necessarily used in the context of stereoscopic display. FVV can be applied in a 2D context. Consequently, the quality assessment protocols differ as they address the quality of the synthesized view in two different contexts (2D visualization and stereoscopic visualization): it is obvious that stereoscopic impairments (such as cardboard effect, crosstalk, keystone, flickering depth, picket-fence, etc., as described in [MIS04]and [BHG08]), which occur in stereoscopic conditions, are not assessed in 2D conditions. Also, distortions detected in 2D conditions may not be perceptible in a stereoscopic context.

- **Assessed factors.** Also, depending on the case of use, except for the conventional image quality, new factors can be considered such as comfort, naturalness, depth rendering.

- **Clear definition of factors** Even though observers' acuity, stereo-acuity and color vision are measured before the tests, and even though experimental trials are included before the sessions, observers are generally non-expert. In addition, they may not be familiar with simulated stereoscopic viewing. There is a risk for erroneous results, due to the novelty of the media display, which may not always be taken into account in these subjective quality assessment methodologies, regarding the asked tasks. The measured factors need to be clearly defined to avoid confusion when rating the different measured factors.

- **Need for no-reference metric.** Another limitation of usual objective metrics concerns the need for non-reference quality metrics. In particular cases of use, like FVV, references are unavailable because the generated viewpoint is virtual. In other words, there is no ground truth allowing a full comparison with the distorted view. Though, assessment tools are required to evaluate the quality of the synthesized views.

The following sections discuss state-of-the-art methods for 3D content subjective and objective assessment.

4.2 Subjective assessment

3D contents and as a consequence virtual views synthesized either from decoded and distorted data, or from original data, need to be assessed. The best assessment tool remains human judgment as long as the right protocol is used. Subjective quality assessment is still delicate while addressing new types of conditions because one has to define the best way to obtain reliable data. Tests are time-consuming and consequently one should issue precise guidelines on how to conduct such experiments to save time and to bound the number of observers. Since stereoscopic vision and DIBR introduce new parameters, the right protocol to assess visual quality with observers is still an open question. The adequate assessment protocol might vary according to the targeted objective that researchers focus on (impact of compression, DIBR techniques comparison, etc.) and the context of use (viewing conditions: 2D or stereoscopic).

In this section, subjective assessment of 3D content is addressed. Since most of the methods proposed for assessing 3D contents rely on methods used for the assessment of conventional 2D images/video sequences, this latter is first discussed. Then, a discussion on the latest proposed assessment methods for 3D content is proposed.

4.2.1 Subjective assessment methodologies

Subjective tests are used to measure image or video quality. The International Telecommunications Union (ITU)[ITU08] is in charge for the recommendations of the most commonly used subjective assessment methods. Several methods exist but there is no 3D-dedicated protocol, because the technology is not mature yet. In the absence of any better 3D-adapted subjective quality assessment methodologies, the evaluation of 3D content is mostly obtained through 2D validated assessment protocols. The available protocols both have drawbacks and advantages and they are usually chosen according to the desired task. The choice for the use of a particular protocols is determined by the distortion range to assess and by the question and the answer targeted by the researcher [Bar09]. The available methodologies differ according to the type of pattern presentation (single-stimulus, double-stimulus, multi-stimulus), the type of voting (quality, impairment, or preference), the voting scale (discrete or continuous), the number of rating points or categories. Fig. 4.1 depicts the proposed classification of subjective methods in [Bar09]. The abbreviations of the methods classified in 4.1 are referenced in Table 4.2.

4.2.2 2D-based subjective quality assessment methodologies for 3D contents

As explained earlier, in the absence of any 3D-adapted methodology, the evaluation of 3D contents mostly relies on 2D validated assessment protocols. Numerous examples can be found in the literature. However, the main criticism of the following examples is that they often assess image quality only.

In [KCF07], Kalva *et al.* have studied the impact of eye dominance and autostereoscopic displays on the quality of 3D video experiences, by using an ACR-like methodology. The participants of these experiences were asked to rate the overall quality of asymmetrically coded stereoscopic video sequences using a subjective evaluation scale from 1 (unacceptable) to 5 (excellent).

In [WYYJ09], Wang *et al.* have built a database of 400 distorted stereoscopic still images assessed by twenty observers using the DSCQS method, with polarized glasses. References are hidden. This term indicates that the original images are presented and assessed by the observer but the stimuli are not explicitly identified as the reference images. However, in these experiments, only the right view is distorted while the left view remains original; eye dominance is thus not taken into account in this work.
In [OS10], Olsson *et al.* have also used the DSCQS method to investigate the relationship between compression schemes and perceived 3D image quality. Two different compression methods are tested (JPEG 2000 and H.264/AVC) and an autostereoscopic display is used for the assessments. From a pair of images (the uncompressed original one and the compressed image) the observers are asked to rate the absolute quality only of both images.

In [CLCM07], Campisi *et al.* have proposed a methodology for subjective assessment of stereoscopic images. These experiments include various compression ratios on stereo-

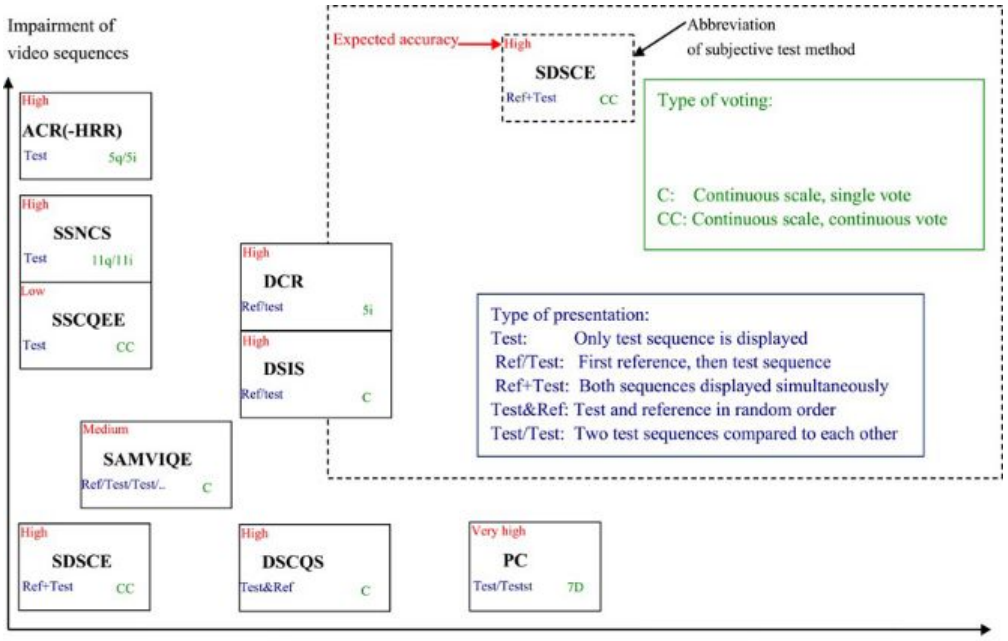


Figure 4.1: *Commonly used subjective test methods, as depicted in [Bar09].*

Abbrev.	Full meaning	Ref.
DSIS	Double Stimulus Impairment Scale	[BT.93]
DSCQS	Double Stimulus Continuous Quality Scale	[BT.93]
SSNCS	Single Stimulus Numerical Categorical Scale	[BT.93]
SSCQE	Single Stimulus Continuous Quality Evaluation	[BT.93]
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation	[BT.93]
ACR	Absolute Category Rating	[ITU08]
ACR-HR	Absolute Category Rating with Hidden Reference removal	[ITU08]
DCR	Degradation Category Rating	[ITU08]
PC	Pair Comparison	[ITU08]
SAMVIQ	Subjective assessment Methodology for Video Quality	[ITU08]

Figure 4.2: *Overview of subjective test methods.*

scopic still images. The tests follow SAMVIQ protocol, with active liquid crystal shutter glasses. SAMVIQ method is known for its ability to discriminate similar levels of quality. Moreover, with a random access process to display the contents, observers are allowed to start and stop the evaluation, to modify their vote or to repeat the display of a media. In this case, references are also hidden and rated by the observers. In spite of its accuracy, it is a time-consuming method.

In [ZW09], Zhu *et al.* used a DSIS method in order to validate the reliability of their proposed perceptual metric for stereoscopic video sequences. Similarly, in the Call for Proposals on 3D Video Coding (3DVC) technology of MPEG [MPE11], the testing method is also DSIS, and with naive observers. In this method, observers are presented two stimuli sequentially. Afterwards, the observers are allowed to rate the impairments for a few seconds only. In [ZW09], the observers rate the quality by answering the question "how close it visually resembles the original reference".

4.2.3 Trends (towards 3D adapted protocols)

Defining a new subjective video quality assessment framework is a tough task, given the new complexity involved in 3D media. The difficulty of 3D-image quality evaluation, compared to 2D conventional images, is now better considered. Seuntiens [Seu06] introduced new parameters to be assessed in addition to image quality: naturalness, presence and visual experience. Thus, a multi-dimensional quality indicator may allow a reliable assessment of 3D-TV media. Yamagashi *et al.* recently studied the relationships between image quality, naturalness and depth in [YKOH11]. ITU-R BT. 1438 recommendation [ITU00] describes subjective assessment of stereoscopic television pictures and the methods are described in [BT.93]. New protocols are under investigation and the relevant trends are discussed hereby.

Chen *et al.*[CFBLC10] revisited the question of subjective video quality assessment protocols for 3D-TV. This work points out the complexity of 3D media quality assessment. Chen *et al.* proposed to reconsider several conditions in this context, such as the viewing conditions (viewing distance, monitor resolution), the test material (depth rendering according to the chosen 3D display), viewing duration, etc. In the following, some of the requirements proposed by Chen *et al.* in [CFBLC10] are mentioned:

- General viewing conditions: first the luminance and contrast ratio is considered, because of the crosstalk involved by 3D-TV screens, and because of the used glasses (both active and polarized glasses cause reduction of luminance). Second, the resolution of depth has to be defined. Third, the viewing distance recommended by ITU standards may differ according to the used 3D display. Moreover, as the authors of the study claim, depth perception should be considered as a new parameter to evaluate the Preferred Viewing Distance, such as human visual acuity or picture resolution.

- Source signals: the video format issue is mentioned. It refers to the numerous 3D representations (namely LDV, MVD, or 2D+Z) whose reconstruction or conversion lead to different types of artifacts.

- Test methods: as mentioned previously, new aspects have to be considered such as naturalness, presence, visual experience and visual comfort as well. The latter refers to the visual fatigue that should be measured to help in a standardization process.

- Observers: first an adapted protocol should involve the measurement of viewers' stereopsis ability. Second, the authors of [CFBLC10] mention that the required number of participants may differ in 2D and in 3D. So, further experiments should define this number.
- Test duration and results analysis: the duration of the test is still to be determined, taking into account visual comfort. Analysis of the results refers to the definition of a criterion for the rejection of incoherent viewer votes and also to the analysis of the assessed parameters (depth, image quality, etc.)

In [AHH+10], Aflaki *et al.* have included the separate evaluation of three criteria: general image quality, naturalness and perceived depth. The discrete quality scale was ranged from -3 to 3 (-3 meant "very bad" or "not natural", 0 meant "mediocre" and 3 was "very good" or "very natural"). In this attempt to measure multi-modes brought by stereoscopic vision, naive participants with no experience on stereoscopic video rate asymmetrically coded video sequences through a polarizing screen. The originality of this work come from the fact that the proposed methodology includes the evaluation of three different factors, and not only the image quality. However, in the paper, the analysis of the separate contribution of these three criteria is unclear.

In [OEK11], Ozbek *et al.* have proposed an interactive quality assessment method for stereoscopic video sequences, namely the Subjective Evaluation of Stereo VIdeo Quality (SESVIQ). It is based on the SAMVIQ method. Observers vote with a slider whose values ranging from 0 to 100 are grouped in five qualitative categories (bad, poor, fair, good and excellent). In this methodology, it is worth noting that three modes are assessed simultaneously, within the same score: the observers are ask to vote considering perceived depth, sharpness and naturalness. As in the SAMVIQ method, observers can repeat the display and compare the impaired sequences as long as they need. The main criticism regarding the proposed protocol is that the separate contribution of these three criteria is not clear. Moreover, as SAMVIQ method, in spite of its accuracy, SESVIQ is time-consuming. In [CJB+12], Chen *et al.* have also used a SAMVIQ methodology. In that paper, the authors consider the visual experience as a linear combination of visual comfort, perceived depth and image quality experienced by the observer. The different factors are separately rated: depth quantity is estimated by the observer through a numerical scale ranging from 0 to 100. The other criteria were estimated through a qualitative scale ("excellent, good, fair, poor, bad". This study revealed the contributions of the considered factors and led to the conclusion that visual experience is a combination of image quality (34%), depth rendering (27%) and visual comfort (40%).

These studies show the need for the consideration of new modes when assessing 3D contents. In consequence, objective metrics also require to be addressed in the context of 3D contents assessment, since they are meant to predict human judgment whose complexity in 3D seem to be different than in 2D. This is the subject of the next section.

4.3 Objective assessment

The need for better adapted tools to correctly assess the quality of 3D contents is crucial. Indeed, the performances of any new system, such as synthesis algorithms or coding methods, need to be determined in order to make the best technology choices. The latest proposed metrics in the literature, do not always consider the same viewing conditions.

For instance, among the proposed metrics, many of them target stereoscopic video, but only a few of them target views synthesized from DIBR in 2D viewing conditions.

Most of the proposed metrics assessing 3D media, are inspired from 2D quality metrics. It should be noted that experimental protocols validating the proposed metrics often involve depth and/or color compression, different 3D displays, and different 3D representations (2D+Z, stereoscopic video, MVD, etc.). Experimental protocols often assess at the same time both compression distortion and synthesis distortion, without distinction. This is problematic because there may be a combination of artifacts from various sources (compression and synthesis) whose effects are not clearly specified and assessed.

The objective metrics can be classified in three different categories of methods according to the availability of the reference image [CSRK11]: full reference methods (FR), reduced reference (RR), and no-reference (NR). Most of the existing metrics rely on FR methods which require references images. RR methods require only elements of the reference images. NR methods do not require any reference images. NR methods mostly rely on Human Visual System (HVS) models to predict the human opinion of the quality. Also, a prior knowledge on the expected artifacts highly improves the design of such methods.

In the following, we present the current trends regarding new objective metrics for 3D media assessment, by distinguishing whether or not they make use of depth data in the quality score computation .

4.3.1 2D-like metrics

In this section, we mention recent studies addressing the issue of objectively assessing 3D contents, and relying on 2D-like metrics. Before presenting these studies, two famous 2D objective tools need to be presented, because of their popularity. These two metrics are known as the Peak-Signal-to-Noise-Ratio (PSNR) and the Single-scale Structural SIMilarity (SSIM) [WBSS04]. They are often used for image quality assessment because of their easiness of implementation.

PSNR measures the signal fidelity of a distorted image compared to a reference. It is based on the measure of the Mean Squared Error (MSE). The MSE between two pictures I and $\tilde{I}$ is defined as follows:

$$MSE = \frac{1}{XY} \sum_{l} \sum_{c} [I(l,c) - \tilde{I}(l,c)]^2 \tag{4.1}$$

where $X \times Y$ is the size of one image, $I(l,c)$ is the value of one pixel in I. The PSNR in decibels is defined as:

$$PSNR = 10log_{10}(\frac{m^2}{MSE}) \tag{4.2}$$

where m is the maximum value that a pixel can take (255 for 8-bit images). Because of the pixel-based approach of such a method, the amount of distorted pixels is summed, but their perceptual impact on the quality is not considered: PSNR does not take into account the visual masking phenomenon. Thus, even if an error is not perceptible, it contributes to the decrease of the quality score. Studies showed that in the case of synthesized views,

PSNR is not reliable, especially when comparing two images with low PSNR scores. PSNR cannot be used in very different scenarios as explained in [ECW04].

SSIM combines image structural information: mean, variance, covariance of pixels, for a single local patch. The block size depends on the viewer's distance from the screen. The SSIM score between two signals x and y is defined as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.3}$$

where, if the two signals x and y contain N samples, the statistical features are:

- $\mu_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $\mu_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
- $\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- $\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$
- $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$
- and the constants: $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$. L is the dynamic range of the pixel values (for 8-bit images, L=255), $K_1 = 0.01$, and $K_2 = 0.03$.

A low variation of the SSIM measure can lead to an important error of MOS prediction. Because of the block based approach of SSIM, it may not be appropriate for the case of 3D content, depending on the case of use.

Despite their limitations, PSNR and SSIM are widely known and used for the assessment of the synthesized views. In the following, recent studies targeting the assessment of 3D contents are presented. Some of them rely or are meant to be extensions of PSNR or SSIM.

The method proposed in [BGE+06] is a full-reference stereoscopic video quality metric combining the evaluation of the monoscopic quality, and the stereoscopic quality of the pair. Inputs are the two views of the stereopair. So-called cyclopean images (combination of the two views into a single global image) are formed from the reference pair, and from the distorted pair. Then, their perceptual similarity is assessed relying on SSIM. Disparities of each pair are analyzed through a block matching algorithm. Correlations with human quality scores (with DSCQS method) are not provided in this paper. However, it is observed that the proposed measure follow the subjective opinion better than SNR (Signal to Noise Ratio), when assessing encoded sequences, or blurred or noised images. This metric does not target the DIBR-related artifacts, but the stereoscopic viewing case only.

Perceptual Quality Metric (PQM) [JMFK10] was proposed by Joveluro *et al.*. Although the authors assess the quality of decoded 3D data (2D+Z), the metric is applied on left and right views synthesized with a DIBR algorithm (namely [Feh04]). Thus, this method may also be applied for synthesized views. The quality score is a weighted function of the contrast distortion and the luminance difference between both reference and distorted color views. The method can thus be classified as HVS-based. The method is sensitive to slight changes in image degradation and error quantification. In [JMFK10] PQM method

performances are validated by evaluating views synthesized from compressed data (both color and depth data are encoded at different bit-rates). Subjective scores are obtained by a SAMVIQ test, on a 3D 42-inch Philips multi-view auto-stereoscopic display. Note that compression, synthesis and factors inherent to the display are assessed at the same time without distinction in the experiments.

Zhao and Yu [ZY10] proposed an FR metric, Peak Signal to Perceptible Temporal Noise Ratio. This metric evaluates the quality of synthesized sequences by measuring the perceptible temporal noise within these impaired sequences.

Conze *et al.*[CRM12] proposed a full-reference objective quality assessment metric that targets artifacts related to view synthesis. More precisely, their method relies on the observation that thin objects, object borders, transparency, variations of illumination or color differences between left and right views, periodic objects are the most critical elements to be rendered through DIBR. Their method is known as the View Synthesis Quality Assessment (VSQA) and is defined as an extension of any existing 2D image quality. In [CRM12], VSQA is used as an extension of SSIM [WBSS04]. VSQA considers features of the spatial environment and the complexity in terms of textures, the diversity of gradient orientations and the presence of high contrast of the synthesized views.

4.3.2 Depth-aided methods

In this section, we present recent studies relying on depth information for the assessment of 3D contents.

Sazzad *et al.* [SYKH09] have proposed a no-reference perceptual quality metric for stereoscopic images based on segmented local features of artifacts and disparity. Detection of local features such as edges, flat and texture blocks, planar and non-planar blocks are included in the method. Disparities are also taken into account in the analysis. The results show that the model performs quite well over a wide range of stereo image contents and distortion levels. However the main criticism to address regards the fact that only JPEG-related artifacts (DCT and blockiness) are targeted by this method.

In [BLCCC09], Benoit *et al.* proposed a quality metric for the assessment of stereopairs using fusion of 2D quality metrics and depth information. Well-known 2D metrics, either SSIM[WBSS04] or C4[CLCB03], are applied separately on each image (left and right view) and the scores are combined to obtain one overall score for the given stereopair. By taking into account the stereo-disparity in their measure, Benoit *et al.* point out the fact that 2D metrics have limitations when assessing stereoscopic image quality, since SSIM is enhanced when adding the disparity distortion contribution. You *et al.*, in [YXPW10] reach the same conclusion regarding the use of disparity in the quality score of stereoscopic data.

Ekmekcioglu *et al.*[EWDS+10] have proposed a depth-based perceptual quality metric. It is a tool that can be applied to PSNR or SSIM. The method uses a weighting function based on depth data at the target viewpoint, and a temporal consistency function to take the motion activity into account. The final score includes a factor that considers non-moving background objects during view synthesis. Inputs of the method are the original depth map (uncompressed), the original color view (originally acquired, uncompressed) and the synthesized view. Validation of the performances is achieved by synthesizing

different viewpoints from distorted data: color views suffer two levels of quantization distortion; depth data suffer four different types of distortion (quantization, low pass filtering, borders shifting, and artificial local spot errors in certain regions). The study [EWDS+10] shows that the proposed method enhances the correlation of PSNR and SSIM to subjective scores.

Yasakethu *et al.*[YWDS+11] proposed an adapted VQM for measuring 3D Video quality. It combines 2D color information quality and depth information quality. Depth quality measurement includes an analysis of the depth planes. The final depth quality measure combines 1) the measure of distortion of the relative distance within each depth plane, 2) the measure of the consistency of each depth plane, and 3) the structural error of the depth. The color quality is based on the VQM score. In [YWDS+11], the metric is evaluated through left and right views (rendered from 2D+Z encoded data), and compared to subjective scores obtained by using an auto-stereoscopic display. Results show higher correlation scores with MOS than simple VQM.

Solh *et al.*[SAB11] have introduced the 3D Video Quality Measure (3VQM) to predict the quality of views synthesized from DIBR algorithms. The method analyses the quality of the depth map against an ideal depth map. Three different analyses lead to three distortion measures: spatial outliers, temporal outliers, and temporal inconsistencies. These measures are combined to provide the final quality score. To validate the method, subjective tests were run in stereoscopic conditions. Stereoscopic pairs include views synthesized from depth map and colored video compression, depth from stereo matching, depth from 2D to 3D conversion. Results show accurate and consistent scores compared to subjective assessments.

In [JQL09], Jiangbo *et al.* have proposed an interpolation quality metric including a image interpolation algorithm to detect color "bleeding artifacts", related to discovered areas. Depth discontinuities are also detected to predict the disparity jumps greater than a threshold. The results are encouraging but the paper does not include comparisons with other known metrics.

4.4 Conclusion

This chapter proposed a review on both subjective quality assessment protocols and objective quality assessment methods used in the context of MVD. This analysis showed that subjective and objective methods tend to take the added-value of depth more into consideration. This makes the evaluation of depth an additional feature to assess, just like the image quality. Most of the proposed objective metrics still rely on 2D usual methods. New tools focus either on depth structure, or on depth accuracy. Temporal consistency is also taken into account. These new aspects need to be consider in view to the conception of 3D Video processing chain. At the time of writing, the Video Quality Experts Group (VQEG) was investigating a new subjective assessment method for 3D services, within the 3DTV project and the studies on an objective metric for 3D video quality was under way. The discussion of the next part of this thesis will remain on quality assessment, in order to investigate a particular aspect regarding the tools evaluating the quality of synthesized views.

Part II

Visual quality assessment of synthesized views

Probable causes of distortions have been presented in the previous part through the examination of the 3D Video processing chain steps. We can now focus on a specific phase generating visual distortions in the end user 3D media: the view synthesis process. This choice is motivated by the fact that both 3DTV and FTV require the reconstruction of novel virtual viewpoints. Since the virtual viewpoints are actually observed by the users, the quality assessment of these generated views is meaningful. In addition, considering, the fact that the processing chain head end systems also require their performances to be rated, it is primordial to ensure the availability of assessment tools for virtual views. This part addresses this issue.
Chapter 5 explains the principles of view synthesis, through the example of the reference algorithm used in this thesis. This chapter also presents the synthesis process related distortions. Chapter 6 illustrates the complexity of synthesized views assessment thanks to three experiments questioning the reliability of subjective quality evaluation and objective quality evaluation methods commonly used for 2D images/video sequences, in the case of the synthesized views assessment. The experiments include both 2D viewing conditions and stereoscopic viewing conditions. Finally, this chapter also proposes a preliminary study for the objective quality assessment of synthesized views, based on the results of the previous experiments.

CHAPTER 5

View synthesis in 3D video

In the context of this thesis, view synthesis refers to the process using depth maps to generate novel viewpoints of the same scene. This warping process involves issues related to geometry. It induces new artifacts in the reconstructed virtual view. The goal of this chapter is to present the possible sources of distortion deriving from the synthesis process. There is a need for new tools such as objective quality metrics or MVD coding methods whose design can be guided based on the knowledge of these sources of distortion. The chapter is organized in two sections: the introduction of the synthesis principles through the example of a synthesis algorithm, namely View Synthesis Reference Software (VSRS) is presented in a first section (Sec. 5.1) and a discussion on the related artifacts is proposed in a second section (Sec. 5.2).

5.1 View synthesis principles

In order to show a good understanding of the artifacts related to the synthesis process, we describe the principles of this technique in this section. Both 3D-TV and FVV applications require the generation of novel viewpoints. The transmitted texture and depth video sequences are used to generate virtual views with the help of Depth-Image-Based Rendering (DIBR) techniques. The generated views can then be rendered on a conventional display, or a stereoscopic or an autosterescopic display.

Generating a "virtual" view consists in synthesizing a novel view of the scene, from a viewpoint which differs from those captured by the actual cameras, by relying on the available texture and depth data. The conventional 2D color sequences provide the color information, also called texture. The depth data is defined by gray-scales images and is considered as a monochromatic signal. Each pixel of a depth map, gives the distance of the corresponding 3D point from the camera, as explained in section 2.2. Based on projective geometry [HZ03], the 3D representation of a scene can be retrieved from a depth map.

Figure 5.1 illustrates the relationship between a real 3D point of the scene, defined as X and its projections x_1 and x_2 in camera planes of C_1 and C_2 respectively. Points x_1 and x_2 are said to be "correspondent pixels" because they are the projection of the same real 3D point X. Given the depth of X and the cameras' parameters, x_1 and x_2 can be

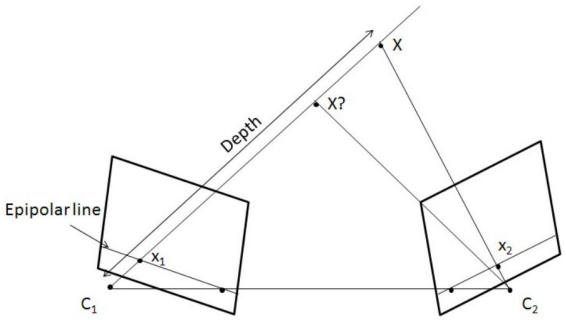


Figure 5.1: *Relationship between image points and real world [LH08].*

determined using projective geometry principles: the geometric transformation from 3D world to the camera plane can be easily performed from depth data, and both intrinsic and extrinsic parameters of known cameras [HZ03].

Along the same principle, 3D points of the real world can be projected onto the image plane of a virtual camera from an arbitrary viewpoint. Since we will use the View Synthesis Reference Software (VSRS) [TFS+08] (version 3.5 is provided by MPEG) as the synthesis algorithm in this thesis, we propose to illustrate the principles of DIBR relying of the fundamentals of this method. The choice for VSRS here and in the rest of this thesis, is motivated by the fact that it is the reference software in MPEG and the proposed coding systems need to be compared based on a common reference. Providing the parameters related to the virtual camera C_v and texture and depth information from two adjacent known views, VSRS is able to generate a novel viewpoint.

Figure 5.2 shows synthesis principles used in VSRS. Let T_{o1} and T_{o2} be the two original texture adjacent views, left and right respectively. Let d_{o1} and d_{o2} be the depth maps of the two adjacent views, T_{o1} and T_{o2} respectively. Depth maps d_{o1} and d_{o2} are warped into the virtual view resulting in two new depth maps referring to the virtual viewpoint: d_{p1} and d_{p2} respectively. Those new maps contain non-valued areas, called holes. They correspond to occluded areas in the reference viewpoint (left or right respectively). The left and right texture images T_{o1} and T_{o2} are projected in the virtual viewpoint according to the new depth maps d_{p1} and d_{p2}, they also contain non-valued areas. The resulting texture images can be denoted as T_{p1} and T_{p2}. The non-valued areas, also called holes, are then filled in by available information from both new texture images. Then the two texture images T_{p1} and T_{p2} are fused into one single image denoted as T_v.
The described process assigns each pixel of the new texture image T_v, a color value according to its corresponding depth. Three cases are considered:

- both depth values for the considered pixel are null: this is a non visible area.
- only one of the two pixels has a depth value: this is an occluded area in one of the reference viewpoints.
- depth values of the pixels in the adjacent views are not null.

This is expressed by:

$$T_v = \begin{cases} 0, & \text{if } (u,v) \text{ is not visible} \\ T_{p1}(u,v), & \text{if } d_{p1}(u,v) \neq 0 \\ & \text{and } d_{p2}(u,v) = 0 \\ T_{p2}(u,v), & \text{if } d_{p1}(u,v) = 0 \\ & \text{and } d_{p2}(u,v) \neq 0 \\ (1-\alpha)T_{p1}(u,v) + \alpha T_{p2}(u,v), & \\ & \text{if } d_{p1}(u,v) \neq 0 \\ & \text{and } d_{p2}(u,v) \neq 0 \end{cases}$$

where (u, v) refers to the coordinates of a pixel of the synthesized view, $d_{p1}(u, v)$ is the depth value of this pixel calculated from camera C_1, $d_{p2}(u, v)$ is the depth value of this pixel calculated from camera C_2, and α is a factor depending on the distance to the virtual viewpoint ($\alpha < 1$). To be more precise, the factor is calculated in a way that the view closer to the virtual view position has a higher weight.

The synthesis process already raises some issues. First, in terms of geometry: regions occluded in both input views and visible in the target view lead to non-valued areas in the texture image. Secondly, errors can occur because pixel coordinates do not locate at an integer position and are usually either interpolated or rounded to the nearest integer position. Inpainting methods [NNKD+10, KNND+10, MSD+08] as well as interpolation filters were developed in order to reduce these synthesis artifacts. However, using such processes lead to new artifacts. This will be discussed in the next section.

5.2 New artifacts

In this section, we first discuss the sources of distortions in synthesized views. Then, we present a classification of commonly observed distortions.

5.2.1 Sources of distortion

The major issue in DIBR consists in filling in the disoccluded regions of the novel viewpoint: when generating a new viewpoint, regions that were not visible in the previous viewpoints, become visible in the new one [Feh04]. However, the appropriate color information related to these discovered regions is often unknown. Inpainting methods that are either extrapolation or interpolation techniques, are meant to fill in the disoccluded regions. Evaluating the impact of such processes on the visual quality is difficult. Distortions from inpainting are specific and dependent on a given hole-filling technique, as observed in [BPLC+11b].

Another noticeable problem with respect to visual quality assessment refers to the numerical rounding of pixel positions when projecting the color information in the target viewpoint (3D warping process): the pixels mapped in the target viewpoint may not be located at an integer position. In this case the position is either rounded to the nearest integer or interpolated. Finally, another source of distortion relies on the depth map uncertainties. Errors in depth maps estimation cause visual distortion in the synthesized views because the color pixels are not correctly mapped onto the new texture image. Similar artifacts occur when depth maps suffer important quantization from compression methods [DSFW10].

5.2.2 Examples of distortions

As explained above, the sources of distortions are various and their visual effects on the synthesized views are perceptible as well in the spatial domain as in the temporal domain. The following statements are based on our observations. In most cases, these artifacts are located around large depth discontinuities, and they are more noticeable in case of high texture contrast between background and foreground. The artifacts are perceptible in monoscopic viewing condition. In stereoscopic viewing condition, depending on the importance of the distortion, either binocular suppression or binocular rivalry can occur,

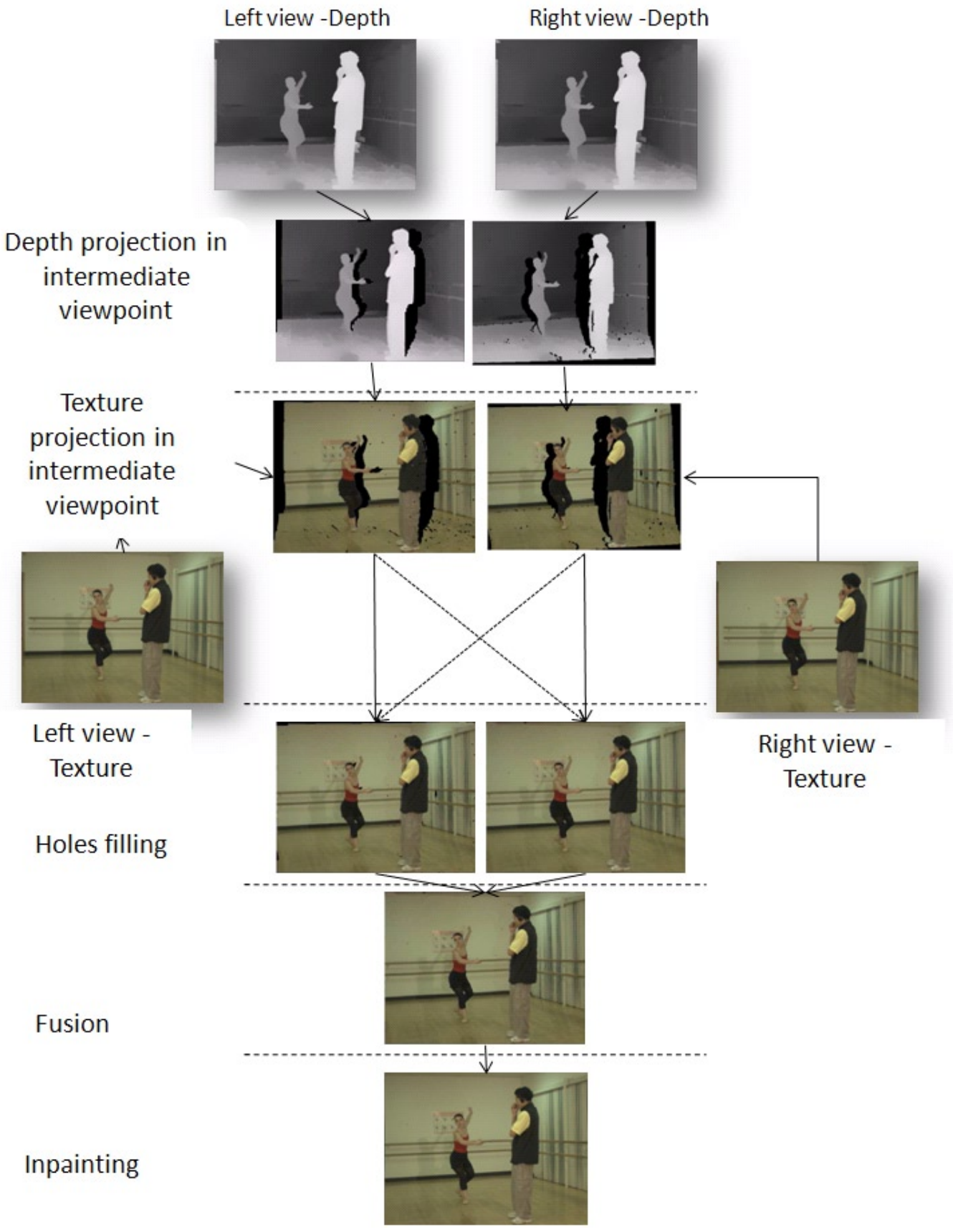


Figure 5.2: *View synthesis with VSRS.*

Figure 5.3: *Shifting/Resizing artifacts. The shape of the leaves, in this figure, is slightly modified (thinner or bigger). The vase is also moved.*

Figure 5.4: *Incorrect rendering of textured areas. An example of texture stretching.*

making the artifacts imperceptible or perceptible, respectively. In this subsection, these different artifacts observed at the end of the synthesis step are addressed. The resulting typical DIBR artifacts are first described. Then, the artifacts observed when synthesizing a novel view from compressed depth data are discussed.

Artifacts related to the synthesis process

Object shifting: a region may be slightly translated or resized, depending on the chosen extrapolation method (if the method chooses to assign the background values to the missing areas, object may be resized), or on the encoding method (in depth data blocking artifacts result in object shifting in the synthesis). Figure 5.3 depicts this type of artifact.

Incorrect rendering of textured areas: inpainting (or hole-filling) methods may fail to fill in complex textured areas. Figure 5.4 depicts this type of artifact.

Blurry regions: This may be due to the inpainting method used to fill in the disoccluded areas. It is more visible around the background/foreground transitions. These remarks are confirmed on Figure 5.5 around the disoccluded areas. Behind the head and around the arms of the chair, thin blurry regions are perceptible.

Flickering: errors occurring randomly in depth data along the sequence imply that color pixels are projected into an erroneous location: some pixels suffer slight changes of depth at successive time instants, which appears as flickers in the resulting synthesized pixels. This can thus be observed when watching a video sequence.

Tiny distortions: in synthesized sequences, a large number of tiny geometric distortions and illumination differences are temporally constant and perceptually invisible. Due

Original frame Synthesized frame

Figure 5.5: *Blurry artifacts (Book Arrival).*

to the rounding decimal point problem mentioned in Section 5.1 and to depth inaccuracy, slight errors may occur when assigning a color value to a pixel in the target viewpoint. This leads to tiny illumination errors, or tiny geometric shifts that may not be perceptible to the human eye. However, pixel-based metrics may penalize these distorted zones.

Artifacts related to the synthesis from compressed depth data

When encoding either depth data or color sequences before performing the synthesis, compression-related artifacts are combined with synthesis artifacts. Artifacts from data compression are generally scattered over the whole image (as described in [YW98]), while artifacts inherent to the synthesis process are mainly located around the disoccluded areas. The combination of both types of distortion, depending on the compression method, relatively degrade the synthesized view. Actually, most of the used compression methods are based on 2D video coding methods, and are thus optimized for the human perception of color. As a result, artifacts occurring especially in depth data induce severe distortions in the synthesized views. In the following, a few examples of such distortions are presented.

Shifting effect: this shifting effect is due to staircase effect or blocking effect in the quantized depth map. This occurs when the DCT based compression method deals with diagonal edges and features. Coarse quantization of blocks containing a diagonal edge results in either a horizontal or vertical reconstruction, depending on its original orientation. In the synthesized views, whole blocks of color image seem to be translated. Figure 5.6 illustrates the distortion. Staircase effect is perceptible in the depth map and it results in geometric distortions of the projected objects: the face and the arms have distorted shapes. The diagonal line in the background is also degraded. The staircase effect modifies the depth plane values of the color pixels, thus objects are consequently wrongly projected during the synthesis.

(a) Original depth frame

(b) Original texture frame

(c) Distorted depth frame

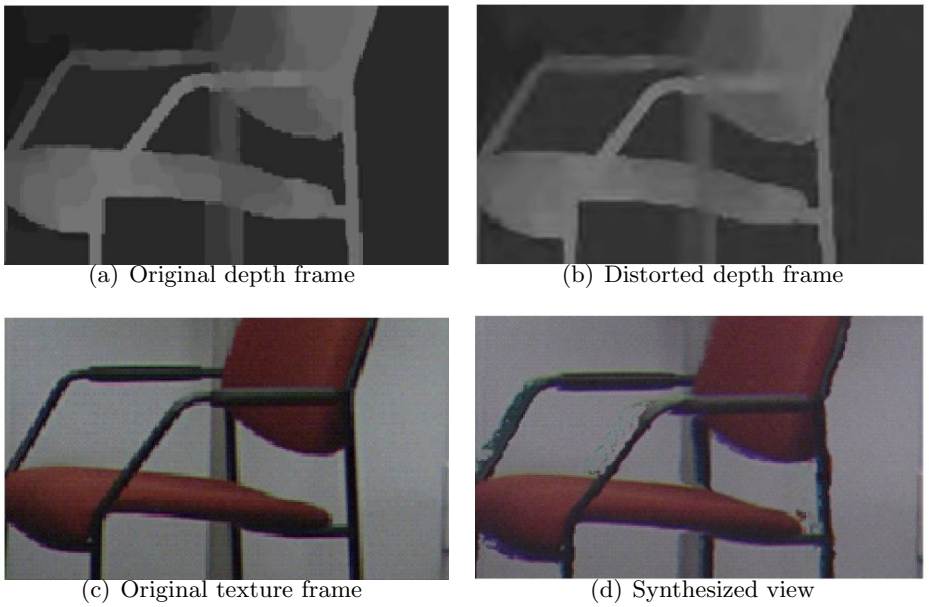
(d) Synthesized view

Figure 5.6: *Shifting effect from depth data compression results in distorted synthesized views (Breakdancers).*

"Crumbling": when artifacts occur in depth data around strong discontinuities, appearing like erosion, the edges of objects appear distorted in the synthesized view. This typically occurs when applying wavelet-based compression on depth data. Figure 5.7 depicts this artifact. It is perceptible around the arms of the chair.

5.3 Conclusion

In this chapter, the principles of view synthesis based on depth images have been introduced, thanks to the example of the DIBR algorithm used in our following experiments, namely VSRS. The presentation of this process highlighted the origins of distortions in synthesized views. This chapter provided examples of typical DIBR related distortions. We observed that visual distortions could be not only the result of the very synthesis process itself (in particular because of the hole filling issue), but as well the result of the combination of compression related artifacts and synthesis related ones. New types of distortions have arisen. Based on these observations, we need to question the reliability of usual quality metrics and usual subjective quality assessment protocols. This will be the subject of the next chapter.



(a) Original depth frame (b) Distorted depth frame

(c) Original texture frame (d) Synthesized view

Figure 5.7: *Crumbling effect in depth data leads to distortions in the synthesized views (Book Arrival).*

CHAPTER 6

Assessment of synthesized views

The previous chapter highlighted the fact that new types of distortion have to be considered due to the synthesis process. This observation justifies the need for the confirmation that usual quality assessment methods are still reliable in the specific case of use of DIBR algorithms.
In this chapter, experiments addressing this verification are presented. They target the quality evaluation of still images and video sequences in the presence of synthesis related artifacts only (i. e. there is no compression related artifacts). Generated virtual views are assessed in monoscopic conditions and in stereoscopic conditions. This study has been conducted with the collaboration of IRCCyN laboratory, at University of Nantes, France, and Fraunhofer Institut for Telecommunications, HHI, Berlin, Germany. It led to the publication of three international conferences papers [BPLC+11a, BKP+11, BMP12b], one journal paper [BPLC+11b] and one book chapter [BLCMP12].
In this chapter, Sec. 6.1 details the motivation for our contribution, through these experiments. The next two sections are dedicated to the presentation and the justification for the choice of the tested subjective quality assessment methodologies (Sec. 6.2) and the objective quality metrics (Sec. 6.3). Then, Sec. 6.4 describes the experimental conditions. The results of each experiments are presented in three different sections (Sec 6.5, Sec. 6.6 and Sec. 6.7) depending on the viewing conditions and on the assessed media.

6.1 Goal of the study

Most of the proposed metrics for assessing 3D media are based on 2D quality metrics. Previous studies ([YHFK08, TGSM08, HWD+09]) already considered the reliability of usual objective metrics. In [YXPW10], You *et al.* studied the assessment of stereoscopic images in stereoscopic conditions with usual 2D image quality metrics, but the distorted pairs did not include any DIBR-related artifacts. In such studies, experimental protocols often involve depth and/or color compression, different 3D displays, and different 3D representations (2D+Z, stereoscopic video, MVD, etc...). In these cases, the quality scores obtained from subjective assessments are compared to the quality scores obtained through objective measurements, in order to find a correlation and validate the objective metric. The experimental protocols often assess both compression distortion and synthesis distortion, at the same time without distinction. This is problematic because there may be a

combination of artifacts from various sources (compression and synthesis) whose effects are neither understood nor assessed.

The experiments presented in this chapter question the reliability of subjective and objective assessment methods when evaluating the quality of synthesized views. Since the goal of these experiments is to question the performances of commonly used quality assessment methods regarding the synthesis related artifacts only, compression related artifacts must not be introduced in the reference data. In other words, the proposed verification protocol will include the evaluation of synthesized views generated from original texture and depth data only, to avoid the mixed assessment of synthesis and compression related artifacts. These experiments involve the use of seven different DIBR algorithms, in order to consider various types of artifacts. Several commonly used objective quality metrics and subjective quality assessment methodologies will evaluate the quality of the synthesized views. A first step consists in determining the reliability of usual methods in monoscopic conditions, because it is a plausible case of use (FVV applications for instance). Then, the stereoscopic conditions are addressed in a second step. In the next sections, we will justify the choice for the selected subjective and objective quality assessment methods. Then, after the detailed presentation of the experimental protocols, the results of different experiments will be discussed.

6.2 Tested subjective assessment methodologies

Based on the review of subjective assessment methodologies proposed in Chapter 4.2.1, we select two subjective quality assessment methodologies. As explained, in the absence of any better 3D-adapted subjective quality assessment methodologies, the evaluation of synthesized views is often obtained through 2D validated assessment protocols. The aim of our experiments is to question the suitability of a selection of subjective quality assessment methods. This selection is based on the comparison of methods in the literature. Considering the aim of the experiments that we proposed, the choice of a subjective quality assessment method should be based on consideration of reliability, accuracy, efficiency and easiness of implementation of the available methods. This section justifies the choice for the subjective quality assessment protocols used in the following experiments.

Brotherton *et al.*[BHHB06] investigated the suitability of ACR and SAMVIQ methods when assessing 2D media. The study showed that ACR method allows more test sequences (at least twice as many) to be presented for assessment compared to the SAMVIQ method. ACR method also proved to be reliable in the test conditions. Rouse *et al.* also studied the trade off of these two methods in [RPLCH10], in the context of high definition still images and video sequences. They concluded that the suitability of the two methods could depend on specific applications.
A study was conducted by Huynh-Thu *et al.* in [HGS+11] to compare different methods according to their different voting scales (5-point discrete, 9-point discrete, 5-point continuous, 11-point continuous scales). The tests were carried out in the context of high-definition video. The results showed that the ACR method produced reliable subjective results, even across different scales.
Considering these analyses of the methods in the literature, we selected the single-stimulus pattern presentation, ACR-HR (with 5 quality categories) and the double-stimulus pattern presentation PC for its accuracy. Both are described and commented in the following.

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 6.1: *Comparison scale for ACR-HR*

Absolute Categorical Rating with Hidden Reference Removal (ACR-HR) methodology consists in presenting test objects (i.e., images or sequences) to observers, one at a time. The objects are rated independently on a category scale. The reference version of each object must be included in the test procedure and rated like any other stimulus. This explains the term used of "hidden reference". From the obtained scores , a differential score (DMOS for Differential Mean Opinion Score) is computed between the mean opinion scores (MOS) of each test object and its associated hidden reference. ITU recommends the 5-level quality scale depicted in Table 6.1. ACR-HR requires many observers to minimize the contextual effects (previously presented stimuli influence the observer's opinion, i.e., presentation order influences opinion ratings). Accuracy increases with the number of participants.

Paired Comparisons (PC) methodology is an assessment protocol in which stimuli are presented by pairs to the observers: it is a double-stimulus method. The observer selects the one out of the pair that best satisfies the specified judgment criterion, i.e. image quality. The results of a paired comparison test are recorded in a matrix: each element corresponds to the frequencies a stimulus is preferred over another stimulus. This data is then converted to scale values using Thurstone-Mosteller's model or Bradley-Terry's [Han01]. It leads to a hypothetical perceptual continuum.
The presented experiments follow Thurstone-Mosteller's model where naive observers are asked to choose the preferred item from one pair. Although the method is known to be highly accurate, it is time consuming since the number of comparisons grows considerably when the number of images to be compared increases.

Discussion on the two methodologies

The differences between ACR-HR and PC are of different types. First, with ACR-HR, even though they may be included in the stimuli, the reference sequences are not identified as such by the observers. Observers assign an absolute grade without any reference. In PC, observers only need to indicate their preference out of a pair of stimuli. Therefore the requested task is different: while observers assess the quality of the stimuli in ACR-HR, they just give their preference in PC.

The quality scale is another issue. ACR-HR scores provide knowledge on the perceived quality level of the stimuli. However, the voting scale is coarse, and because of the single stimulus presentation, observers cannot remember previous stimuli and precisely evaluate small impairments. PC scores (i.e. "preference matrices") are scaled to a hypothetical perceptual continuum. However, it does not provide knowledge on the quality level of the stimuli, but on the order of preference. On the other hand, PC is very well suited for small impairments, thanks to the fact that only two conditions are compared each time. This is

why PC tests are often coupled with ACR-HR tests.

Another aspect concerns the complexity and the feasibility of the test: PC is simple because observers only need to provide preference in each double stimulus. However, when the number of stimuli increase, the test becomes difficult to carry out since the number of comparisons grows according to $\frac{N(N-1)}{2}$, where N is the number of stimuli. In the case of video sequence assessment, a double-stimulus method such as PC involves the use of either one split-screen environment (or two full screens), with the risk of distracting the observer (as explained in [PW03]), or one screen but the length of the test increases as sequences are displayed one after the other. On the other hand, the ease of handling of ACR-HR allows the assessment of a larger number of stimuli but, the results of this assessment are reliable as long as the group of participants is large enough.

6.3 Tested objective metrics

As presented in Chapter 4.3, proposed metrics for 3D media often rely on 2D metrics. Considering the fact that the synthesis process induces artifacts whose influence on objective metrics performances has not been addressed so far, we propose to include a selection of commonly used 2D metrics in our experiments and to validate their reliability in the context of use of DIBR.
The choice of the objective metrics used in these experiments is motivated by their availability. This section presents an overview of the selected metrics. Still-images and video sequences metrics are presented in the following.

All the objective metrics, FR, RR or NR, can be classified according to a different criterion than the requirement of the reference image. As proposed in [Pó8], we use a classification relying on tools used in the methods presented hereafter. Table 6.2 lists a selection of commonly used objective metrics and Figure 6.1 depicts the proposed classification.

Signal-Based methods:

In this category, only PSNR will be mentioned because it is the most widely used method, due to its simplicity. This FR metric has been presented in Chapter 4.3.

Perception-oriented methods:

Considering that signal-based methods are unable to correctly predict the perceived quality, perception-oriented metrics have been introduced. They make use of perceptual criteria such as luminance or contrast distortion.

UQI [WB02] is an FR perception-oriented metric. The quality score is the product of the correlation between the original and the degraded image, a term defining the luminance distortion and a term defining the contrast distortion. The quality score is computed within a sliding window and the final score is defined as the average of all local scores.

IFC [SBdV05] uses a distortion model to evaluate the information shared between the reference image and the degraded image. IFC indicates the image fidelity rather than the distortion. IFC is based on the hypothesis that, given a source channel and a distortion channel, an image is made of multiple independently distorted subbands. The quality score is the sum of the mutual information between the source and the distorted images

	Objective metric	Abbrev.	Tested
Signal-based	Peak Signal to Noise Ratio	PSNR	X
Perception-oriented	Universal Quality Index	UQI	X
	Information Fidelity Criterion	IFC	X
	Video Quality Metric	VQM	X
	Perceptual Video Quality Measure	PVQM	
Structure-based	Single-scale Structural SIMilarity	SSIM	X
	Multi-scale SSIM	MSSIM	X
	Video Structural Similarity Measure	V-SSIM	X
	Motion-based Video Integrity Evaluation	MOVIE	
HVS-based	PSNR- Human Visual System	PSNR-HVS	X
	PSNR-Human Visual System Masking model	PSNR-HVSM	X
	Visual Signal to Noise Ratio	VSNR	X
	Weighted Signal to Noise Ratio	WSNR	X
	Visual Information Fidelity	VIF	X
	Noise Quality Measure	NQM	X
	Moving Pictures Quality Metric	MPQM	

Table 6.2: *Overview of commonly used objective quality metrics*

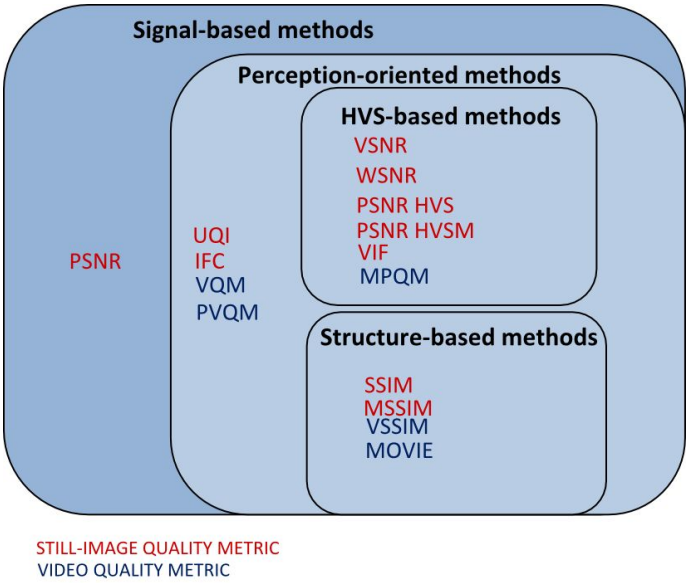


Figure 6.1: *Overview of quality metrics as proposed in [Pó8].*

for all the subbands. It is an FR image quality assessment metric.

VQM was proposed by Pinson and Wolf in [PW04]. It is a RR video metric that measures the perceptual effects of numerous video distortions. It includes a calibration step (to correct spatial/temporal shift, contrast, and brightness according to the reference video sequence) and an analysis of perceptual features. VQM score combines all the perceptual calculated parameters. VQM method is complex but its correlation to subjective scores is good according to [Wan06]. The method is validated in video display conditions.

Perceptual Video Quality Measure (PVQM) [HBL+02] is meant to detect perceptible distortions in video sequences. Various indicators are used. First, an edge-based indicator allows the detection of distorted edges in the images. Second, a motion-based indicator analyzes two successive frames. Third, a color-based indicator detects non-saturated colors. Each indicator is pooled separately across the video and incorporated in a weighting function to obtain the final score. As this FR method was not available, it was not tested in our experiments.

Structure-based methods:

Structure-based methods are perception-oriented metrics that rely on the assumption that human perception is based on the extraction of structural information. Thus, they measure the structural information degradation.

SSIM [WBSS04] was the first method among those of this category. It is considered as an extension of UQI. This FR image quality metric has been presented in Section 4.3. Therefore, many improvements to SSIM were proposed, and adaptations to video assessment were introduced. MSSIM is the average SSIM scores of all patches of the image. V-SSIM [WLB04] is an FR video quality metric which uses structural distortion as an estimate of perceived visual distortion. At patch level, the score is a weighted function of SSIM for the different color components of the image (i.e. luminance and chrominance). At frame level, the score is a weighted function of patches' SSIM scores (the weights depend on the mean value of the luminance in the patch [WLB04]). Finally, at sequence level, VSSIM score is a weighted function of frames' SSIM scores (based on motion). The choice of the weights relies on the assumption that dark regions are less salient. However, this is questionable because the relative luminance may depend on the used screen.

MOVIE[SB10] is an FR video metric that uses several steps before computing the quality score. It includes the decomposition of both reference and distorted video by using a multi-scale spatio-temporal Gabor filter-bank. An SSIM-like method is used for spatial quality analysis. An optical flow calculation is used for motion analysis. Spatial and temporal quality indicators determine the final score.

Human-Visual-System (HVS)-based methods:

HVS-based methods rely on human visual system modeling from psychophysics experiments. Due to the complexity of the human vision, studies are still in progress. HVS-based models are the result of trade-offs between computational feasibility and accuracy of the model. HVS-based models can be classified into two categories: neurobiological models and models based on the psychophysical properties of human vision. The models based on

neurobiology estimate the actual low-level process in human visual system including retina and optical nerve. However, these models are not widely used because of their complexity [BPGA]. Psychophysical HVS-based models are implemented in a sequential process that includes luminance masking, color perception analysis, frequency selection, contrast sensitivity implementation (based on the contrast sensitivity function CSF [YM94]) and modeling of masking and facilitation effects [Win05].

PSNR-HVS [EAP+06], based on PSNR and UQI, takes into account the Human Visual System (HVS) properties such as its sensitivity to contrast change and to low frequency distortions. In [EAP+06], this FR method proved to be correlated to subjective scores, but the performances of the PSNR-HVS method are tested on a variety of distortions specific to 2D image compression which are different from distortions related to DIBR.

PSNR-HVSM [PSE+07] is based on PSNR but takes into account Contrast Sensitivity Function (CSF) and "between-coefficient contrast masking of DCT basis functions". The performances of this FR method are validated considering a set of images containing Gaussian noise or spatially correlated additive Gaussian noise, at different locations (uniformly through the entire image, mostly in regions with a high masking effect or, with a low masking effect).

VSNR [CH07] is also an FR perception-oriented metric: it is based on a visual detection of the distortion criterion, helped by the CSF. VSNR metric is sensitive to geometric distortions such as spatial shifting and rotations, transformations which are typical in DIBR applications.

WSNR that uses a weighting function adapted to HVS denotes a Weighted Signal to Noise Ratio, as applied in [DKG+02]. It is an improvement on PSNR that uses a CSF-based weighting function. So it is also an FR quality metric. However, although WSNR is more accurate by taking into account perceptual properties, the problem remains the accumulation of degradations errors even in non-perceptible areas, like with PSNR method.

NQM was proposed in [DKG+02] as a nonlinear noise quality measure. NQM is an FR quality metric. The appearance of the original and restored images are simulated and the SNR is computed for their difference. The simulation of the appearance of the images as observed by a user is obtained through a nonlinear space-frequency processing based on a modified Peli's contrast pyramid. In particular the CSF is included in the model.

IFC has been improved by the introduction of an HVS model. The FR method is called VIF[SB06]. VIFP is a pixel-based version of VIF. It uses wavelet decomposition and computes the parameters of the distortion models, which enhance the computational complexity. In [SB06], five distortion types are used to validate the performances of the method (JPEG and JPEG 2000 related distortions, white and Gaussian noise over the entire image), which are quite different from the DIBR related artifacts.

MPQM [VLV96] uses an HVS model. In particular it takes into account the masking phenomenon and contrast sensitivity. It has high complexity and its correlation to subjective scores varies according to [Wan]. Since the method is not available, it is not tested in our experiments.

To evaluate the presented metrics, experiments were carried out in monoscopic and stereoscopic viewing conditions. Table 6.2 mentions the metrics actually tested in third column. The next sections will describe the protocols in details and the results of the experiments.

6.4 Experimental framework

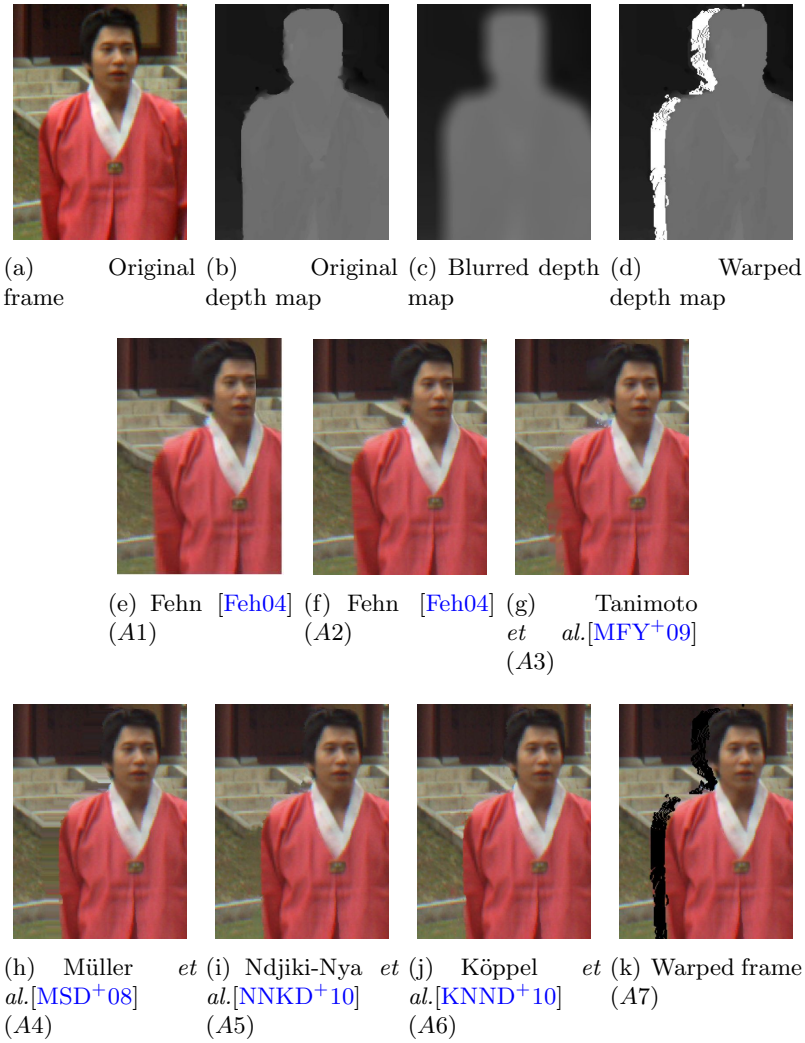
In this section, the common experimental framework of this study is presented. A first subsection presents the tested material and a second subsection addresses the experimental protocols. As expressed previously, the goal of the studies is to evaluate the performances of the objective quality metrics and subjective quality assessment methods that are commonly for 2D media, when dealing with DIBR related artifacts only.

6.4.1 Experimental material

Three different MVD sequences are used in the two studies. The sequences are Book Arrival (1024×768, 16 cameras with 6.5cm spacing), Lovebird1 (1024×768, 12 cameras with 3.5 cm spacing) and Newspaper (1024×768, 9 cameras with 5 cm spacing).
Seven DIBR algorithms processed the three sequences to generate four different viewpoints per sequence. These seven DIBR algorithms are labeled from A1 to A7:

- A1: based on Fehn [Feh04], where the depth map is pre-processed by a low-pass filter. Borders are cropped, and then an interpolation is processed to reach the original size.
- A2: based on Fehn [Feh04]. Borders are inpainted by the method proposed by Telea [Tel04].
- A3: Tanimoto *et al.* [MFY+09]. It is the recently adopted reference software for the experiments in the 3D Video group of MPEG.
- A4: Müller *et al.*[MSD+08] proposed a hole-filling method aided by depth information.
- A5: Ndjiki-Nya *et al.* [NNKD+10]. The hole-filling method is a patch-based texture synthesis.
- A6: Köppel *et al.*[KNND+10] uses depth temporal information to improve the synthesis in the disoccluded areas.
- A7: corresponds to the unfilled sequences (i.e. with holes).

Figure 6.2 gives snapshots of some of the resulted synthesized views.



(a) Original frame (b) Original depth map (c) Blurred depth map (d) Warped depth map

(e) Fehn [Feh04] ($A1$) (f) Fehn [Feh04] ($A2$) (g) Tanimoto *et al.*[MFY+09] ($A3$)

(h) Müller *et al.*[MSD+08] ($A4$) (i) Ndjiki-Nya *et al.*[NNKD+10] ($A5$) (j) Köppel *et al.*[KNND+10] ($A6$) (k) Warped frame ($A7$)

Figure 6.2: *DIBR results for frame 141 of the "Lovebird1" sequence. A baseline of 7cm is used. Camera 8 is used to render camera 6. (a) Original view. (b) Corresponding depth map. (c) Gaussian filtered depth map. (d) Warped original depth map with disocclusions (marked white).(e) "Virtual" camera view obtained with the method proposed by Fehn [Feh04] using the Gaussian filtered depth map shown in (c). (f) based on Fehn [Feh04], borders are inpainted. (g) Result of the hole filling method proposed by Tanimoto et al.[MFY+09]. (h) Result of the hole filling method proposed by Müller et al.[MSD+08]. (i) Result of the hole filling method proposed by Ndjiki-Nya et al.[NNKD+10]. (j) Result of the hole filling method proposed by Köppel et al.[KNND+10]. (k) Corresponding projected "virtual" view with disocclusions (marked black) using the original depth map (b).*

6.4.2 Experimental protocols

Our study can be divided into three different experiments:

- Experiment 1 concerns still synthesized images in monoscopic conditions;
- Experiment 2 concerns synthesized video sequences in monoscopic viewing conditions;
- Experiment 3 concerns still synthesized images in stereoscopic viewing conditions.

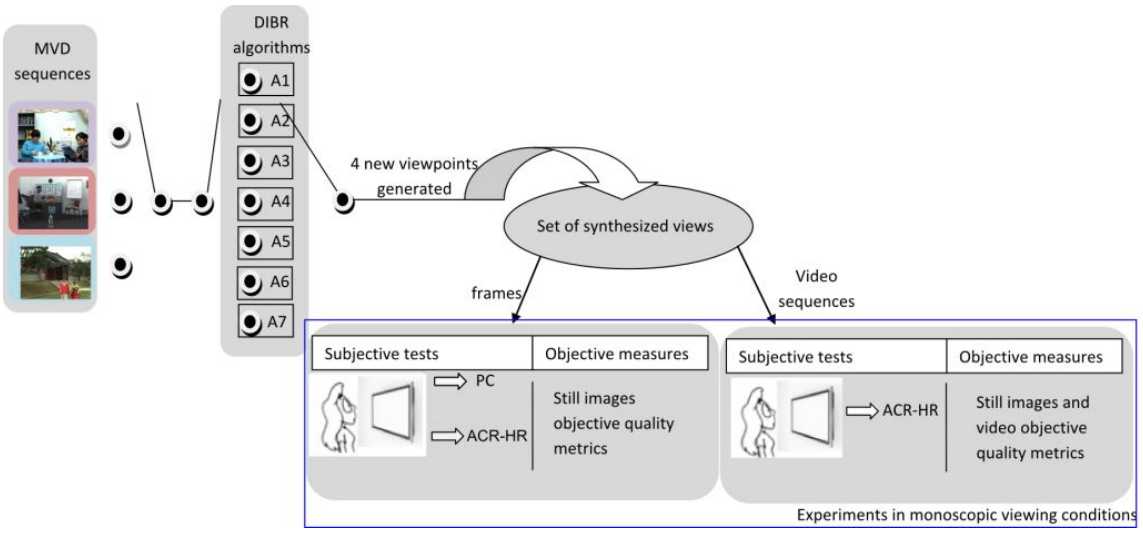


Figure 6.3: *Experimental protocol in monoscopic viewing conditions.*

At the moment of writting this thesis, an experiment regarding the assessment of synthesized video sequences in stereoscopic viewing conditions was expected to be carried. The experimental conditions of each experiment are described in the following.

Experiment 1 and Experiment 2: monoscopic viewing conditions

These two experiments target monoscopic viewing conditions. This viewing case is considered because 2D visualization of synthesized views is possible in FVV applications. A scenario worth studying is when watching a video, the user presses the "pause" button; or the case of 3D display advertisings is also imaginable. These very likely cases are interesting since the image can be subject to meticulous observation. Experiment 1 addresses this scenario. Experiment 2 addresses the evaluation of video sequences. Figure 6.3 depicts the overview of the two experiments.

In both experiments, the suitability of subjective quality assessment methods and the reliability of objective metrics are addressed. Concerning the subjective tests, two sessions were conducted. Forty-three naive observers participated in Experiment 1. The second session addressed the assessment of video sequences: thirty-two naive observers participated in Experiment 2. In both experiments, the subjects were screened for visual acuity. Both ACR-HR and PC were carried out for the still-image context (Experiment 1), but in the case of video sequences (Experiment 2), only an ACR-HR test was conducted. A PC test with video sequences would have required either two screens, or switching between items. In the case of the use of two screens, it involves the risk of missing frames of the tested sequences because one cannot watch two different video sequences simultaneously. In the case of the switch, it would have considerably increased the length of the test.
Table 6.3 summarizes the experimental framework.
The material for both experiments comes from the same set of synthesized views as described in Section 6.4.1. However, in the case of Experiment 1, on still-images, the test images are "key" frames ("keys" were randomly chosen) from the same set of synthesized views. That is to say that for each of the three reference sequences, only one frame was selected out of each synthesized viewpoint, for the tests assessing still-images.
The objective measurements were realized over 84 synthesized views by the means of

		Experiment 1(still-images)	Experiment 2 (video sequences)
Stimuli		Key frames of each synthesized view	Synthesized video sequences
Subjective tests	**No. of participants**	43	32
	Methods	ACR-HR, PC	ACR-HR
Objective measures		All available metrics of MetriX MuX	VQM, VSSIM, Still-image metrics

Table 6.3: *Overview of the experiments*

MetriX MuX Visual Quality Assessment Package [Mux] software, except for two metrics: VQM and VSSIM. VQM was available at [Res]; VSSIM was implemented by our means, with Matlab, according to [WLB04]. The reference was the original acquired image. It should be noted that still image quality metrics used in the study with still images, are also used to assess the visual video sequences quality by applying these metrics on each frame separately and averaging the frames scores.
The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [BT.93]. The stimuli sequences were displayed at the sequence resolution (1024x768) with a grey surrounding to fit the Full HD screen. Objective measurements were obtained by using MetriX MuX Visual Quality Assessment Package [Mux].

Experiment 3: stereoscopic viewing conditions

Only one experiment was conducted in stereoscopic viewing conditions and it targeted the evaluation of still-images. Figure 6.4 depicts the overview of the experimental protocol.
The material comes from the same set of synthesized views as described in Section 6.4.1. The stereopairs consist of two stereo-compliant views. One view is the original acquired frame and the other is a synthesized frame. All the synthesized frames used in this experiments are exactly the same as those used in Experiment 1 (in monoscopic viewing conditions, with still-images).
As in Experiment 1 and Experiment 2, the suitability of subjective quality assessment methods and the reliability of objective metrics are addressed in Experiment 3.
Only one session was conducted using ACR-HR methodology with 25 naive observers. The observers were screnned for visual acuity and for stereo depth perception. The stimuli were displayed on an Acer GD245HQ screen, with NVIDIA 3D Vision Controller. The stimuli sequences were displayed at the sequence resolution (1024x768) with a grey surrounding to fit the Full HD screen.

The objective measurements were realized over 84 synthesized views of the 84 tested stereopairs by the means of MetriX MuX Visual Quality Assessment Package [Mux] software. The reference was the original acquired image.
Table 6.4 summarizes the experimental framework.

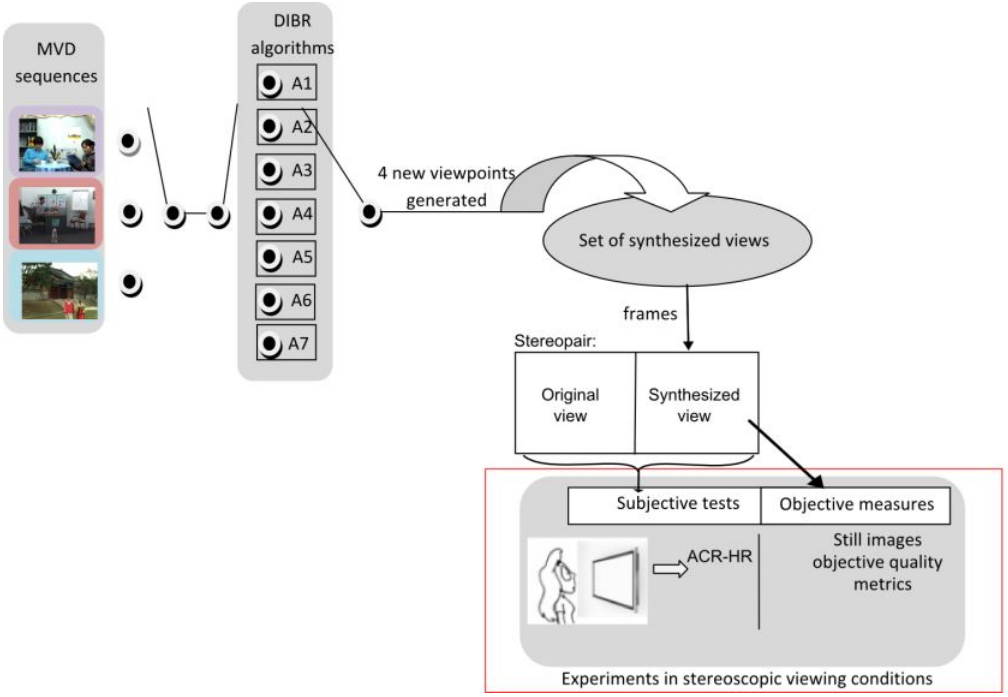


Figure 6.4: *Experimental protocol in stereoscopic viewing conditions.*

		Experiment 3 (stereoscopic still-images)
Stimuli		Stereopairs made up of key frames of each synthesized view for left or right view, and original acquired frame for left or right view
Subjective tests	**No. of participants**	25
	Method	ACR-HR
Objective measures		All available metrics of MetriX MuX

Table 6.4: *Overview of the experiments in stereoscopic viewing*

6.5 Experiment 1: still images in monoscopic conditions

This section addresses the context of still images. These experiments are meant to determine:

- whether the subjective protocols are appropriate for the assessment of different DIBR;
- the number of participants required in such a subjective assessment test;
- whether the results of the subjective assessments are consistent with the objective evaluations.

The following subsections describe the results of Experiment 1. The first part addresses the results of the subjective assessments and the second part presents the results of the objective evaluations.

6.5.1 Subjective tests

The seven DIBR algorithms are ranked according to the obtained ACR-HR and PC scores, as depicted in Table 6.5. For the ACR-HR test, the first line gives the DMOS scores obtained through the MOS scores. For the PC test, the first line gives the hypothetical MOS scores obtained through the comparisons. For both tests, the second line gives the ranking of the algorithms, obtained through the first line. This table indicates that the rankings obtained by both testing methods are consistent except for the ranking of $A2$ and $A6$.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
ACR-HR	3.572	3.308	3.145	3.401	3.496	3.32	2.277
Rank order	1	5	6	3	2	4	7
PC	1.776	0.779	0.338	0.825	1.745	0.61	-2.943
Rank order	1	4	6	3	2	5	7

Table 6.5: *Rankings of algorithms according to subjective scores (still images).*

In Table 6.5, although the algorithms can be ranked from the scaled scores, there is no information concerning the statistical significance of the quality difference of two stimuli (one preferred to another one). Therefore statistical analyses were conducted over the subjective measurements: a Student's t-test was performed over ACR-HR scores, and over PC scores, for each algorithm. This provides knowledge on the statistical equivalence of the algorithms. Table 6.6 and Table 6.7 show the results of the statistical tests over ACR-HR and PC values respectively. In both tables, the number in parentheses indicates the minimum required number of observers that allows statistical distinction (VQEG recommends 24 participants as a minimum in the Multimedia test Plan [Gro], values in bold are higher than 24 in the table).

A first analysis of these two tables indicates that the PC method leads to clear-cut decisions, compared to the ACR-HR method: indeed, the distributions of the algorithms are statistically distinguished with less than 24 participants in 17 cases with PC (only 11 cases with ACR-HR). In one case (between A2 and A5), less than 24 participants are required with PC whereas more than 43 participants are required to establish the statistical difference with ACR-HR. The latter case can be explained by the fact that the

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑**(32)**	↑(<24)	↑**(32)**	**o (>43)**	↑**(30)**	↑(<24)
$A2$	↓**(32)**		↑(<24)	**o (>43)**	**o (>43)**	**o (>43)**	↑(<24)
$A3$	↓(<24)	↓(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓**(32)**	**o(>43)**	↑(<24)		**o(>43)**	**o(>43)**	↑(<24)
$A5$	**o(>43)**	**o(>43)**	↑(<24)	**o(>43)**		↑**(28)**	↑(<24)
$A6$	↓**(30)**	**o(>43)**	↑(<24)	**o (>43)**	↓**(28)**		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓ (<24)	↓(<24)	↓(<24)	

Table 6.6: *Results of Student's t-test with ACR-HR results (still images).* *Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.*

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)
$A2$	↓(<24)		↑**(28)**	o(<24)	↓(<24)	**o(>43)**	↑(<24)
$A3$	↓(<24)	↓**(28)**		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓(<24)	**o(>43)**	↑(<24)		↓(<24)	↑**(43)**	↑(<24)
$A5$	↓(<24)	↑(<24)	↑(<24)	↑(<24)		↑(<24)	↑(<24)
$A6$	↓(<24)	**o(>43)**	↑(<24)	↓**(<43)**	↓(<24)		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	

Table 6.7: *Results of Student's t-test with Paired comparisons results (still images).* *Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "less than 24" observers participate.*

visual quality of the synthesized images may be perceived very similar for non-expert observers. That is to say that the distortions, though different from an algorithm to another, are difficult to assess. The absolute rating task is more delicate for observers than the comparison task. These results indicate that it seems more difficult to assess the quality of synthesized views than to assess the quality of degraded images in other contexts (for instance, quality assessment of images distorted through compression). The results with the ACR-HR method, in Table 6, confirm this observation: in most of the cases, more than 24 participants (or even more than 43) are required to distinguish the classes (Note that A7 is the synthesis with holes around the disoccluded areas).

However, as seen with rankings results above, methodologies give consistent results: when the distinction between algorithms is clear, the ranking is the same with either methodology.

Finally, these experiments show that fewer participants are required for a PC test than for an ACR-HR test. However, as stated before, PC tests, while efficient, are feasible only with a limited number of items to be compared. Another problem, pointed out by these experiments, concerns the assessment of similar items: with both methods, 43 participants were not always sufficient to obtain a clear and reliable decision. Results suggest that observers had difficulties assessing the different types of artifacts.

As a conclusion, this first analysis reveals that more than 24 participants may be necessary for still image quality assessment. Regarding the evaluation of PC and ACR-HR methods, PC gives clear-cut decisions, due to the mode of assessment (preference) while algorithm's statistical distinction with ACR-HR is slightly less accurate. With ACR-HR, the task is not easy for the observers because the impairments among the tested images are small, though each DIBR induces specific artifacts. Thus, this aspect should be taken into account when evaluating the performances of different DIBR algorithms with this methodology.

However, ACR-HR and PC are complementary: when assessing similar items, like in this case study, PC can provide a ranking, while ACR-HR gives the overall perceptual quality of the stimuli.

6.5.2 Objective measurements

The results of this subsection concern the measurements conducted over the same selected "key" frames as those in the previously described subjective test. The objective is to determine the consistency between the subjective assessments and the objective evaluations, and the most consistent objective metric.
The first step consists in comparing the objective scores with the subjective scores (previously presented). The consistency between objective and subjective measures is evaluated by calculating the Pearson linear correlation coefficients (PLCC) for the whole fitted measured points.
The fitting is computed according to the subjective scores using the logistic function recommended by the Video Quality Expert Group (VQEG) Phase I FR-TV [G+00]. The logistic function we used allows the computation of the predicted mean opinion score of a stimuli and it is defined as follows:

$$DMOS_p = a.score^3 + b.score^2 + c.score + d \tag{6.1}$$

where *score* is the obtained score from the objective metric and $\{a, b, c, d\}$ are the parameters of the cubic function. They are obtained through the regression step to minimize the difference between $DMOS$ and $DMOS_p$. The Pearson linear correlation coefficients are then computed though:

$$PLCC = \frac{\sum_{i=1}^{N} \left(DMOS_i - \overline{DMOS}\right)\left(DMOS_{p_i} - \overline{DMOS_p}\right)}{\sqrt{\sum_{i=1}^{N} \left(DMOS_i - \overline{DMOS}\right)^2}\sqrt{\left(DMOS_{p_i} - \overline{DMOS_p}\right)^2}} \tag{6.2}$$

where $\overline{DMOS}$ and $\overline{DMOS_p}$ are the average of $DMOS$ and $DMOS_p$ over the N stimuli.

The coefficients are presented in Table 6.8. In the results of our test, the tested metrics were not correlated to human judgment. This reveals that the objective tested metrics do not reliably predict human appreciation in the case of synthesized views, even though efficiency has been shown for the quality assessment of 2D conventional media.
The whole set of objective metrics gives the same trends. Table 6.10 provides correlation coefficients between obtained objective scores. It reveals that they are highly correlated. This table shows that the behavior of the tested metrics is the same when assessing images containing DIBR related artifacts. Thus, they have the same response when assessing DIBR related artifacts. Note the high correlation scores between pixel-based and more perception-oriented metrics such as PSNR and SSIM (83.9%).
Since it is established in [EPLC+11, HB06] that correlation is different from agreement (as illustrated in Figure 6.5), we check the agreement of the tested metrics by comparing the ranks assigned to the algorithms. Table 6.9 presents the rankings of the algorithms obtained from the objective scores. Rankings from subjective scores are mentioned for comparison. They present a noticeable difference concerning the ranking order of A1: ranked as the best algorithm out of the seven by the subjective scores, it is ranked as the worst by the objective metrics. Another comment refers to the assessment of A6: often regarded as the best algorithm with the objective metrics, it is not the case with the subjective tests. The ensuing assumption is that objective metrics detect and penalize non-annoying artifacts.

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR HVSM	PSNR HVS
CCDMOS	50.0	40.4	57.4	35.0	31.3	22.2	19.1	22.3	57.2	47.7	44.3	42.7
CCPC	36.4	23.1	43.2	16.8	18.2	18.3	24.8	17.7	37.9	33.9	37.5	36.6

Table 6.8: *Pearson correlation coefficients between DMOS and objective scores in percentage (still images).*

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.572	3.308	3.145	3.401	3.496	3.320	2.277
Rank order	1	5	6	3	2	4	7
PC	1.776	0.779	0.338	0.825	1.745	0.610	-2.943
Rank order	1	4	6	3	2	5	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Rank order	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Rank order	7	4	6	1	2	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Rank order	7	4	6	1	2	3	5
VSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Rank order	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Rank order	7	5	6	2	3	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Rank order	7	5	6	2	3	1	4
UQI	0.352	0.672	0.589	0.606	0.605	0.606	0.673
Rank order	7	2	6	3	5	4	1
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Rank order	7	5	6	2	3	1	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Rank order	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Rank order	7	4	5	3	2	1	6
PSNR HSVM	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Rank order	7	4	5	3	2	1	6
PSNR HSV	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Rank order	7	4	5	3	2	1	6

Table 6.9: *Rankings according to measurements (still images).*

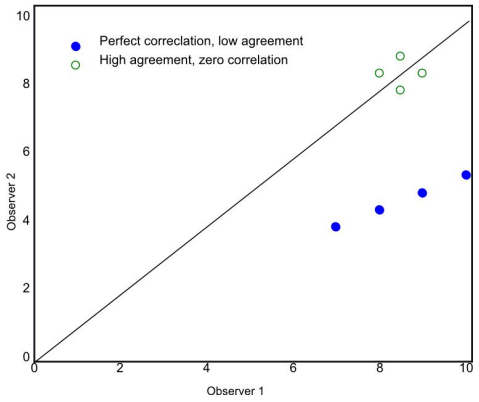


Figure 6.5: *Difference between correlation and agreement [EPLC+11].*

6.5.3 Conclusion

This experiment, involving the assessment of still synthesized images in monoscopic viewing conditions, showed that Paired comparisons and ACR results highly correlate. But statistical analyses showed that fewer observers were required for Paired comparisons tests

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR hsvm	PSNR hsv
PSNR		83.9	79.6	87.3	77.0	70.6	53.6	71.6	95.2	98.2	99.2	99.0
SSIM	83.9		96.7	93.9	93.4	92.4	81.5	92.9	84.9	83.7	83.2	83.5
MSSIM	79.6	96.7		89.7	88.8	90.2	86.3	89.4	85.6	81.1	77.9	78.3
VSNR	87.3	93.9	89.7		87.9	83.3	71.9	84.0	85.3	85.5	86.1	85.8
VIF	77.0	93.4	88.8	87.9		97.5	75.2	98.7	74.4	78.1	79.4	80.2
VIFP	70.6	92.4	90.2	83.3	97.5		85.9	99.2	73.6	75.0	72.2	72.9
UQI	53.6	81.5	86.3	71.9	75.2	85.9		81.9	70.2	61.8	50.9	50.8
IFC	71.6	92.9	89.4	84.0	98.7	99.2	81.9		72.8	74.4	73.5	74.4
NQM	95.2	84.9	85.6	85.3	74.4	73.6	70.2	72.8		97.1	92.3	91.8
WSNR	98.2	83.7	81.1	85.5	78.1	75.0	61.8	74.4	97.1		97.4	97.1
PSNR hsvm	99.2	83.2	77.9	86.1	79.4	72.2	50.9	73.5	92.3	97.4		99.9
PSNR hsv	99.0	83.5	78.3	85.8	80.2	72.9	50.8	74.4	91.8	97.1	99.9	

Table 6.10: *Correlation coefficients between objective metrics in percentage (still images).*

to establish the algorithms distinctions. However, this is a time-consuming method, often avoided by researchers. Moreover, when the number of items to be compared is high, the test is hardly feasible. Concerning the objective metrics, the results showed that usual objective assessments do not correlate with subjective assessments. Rankings of algorithms from objective metrics are not reliable, considering the differences with the obtained subjective results. The presented experiments revealed that using only the objective metrics seems not sufficient for assessing virtual synthesized views, though they give information on the presence of errors.

6.6 Experiment 2: video sequences in monoscopic conditions

This section addresses the evaluation of video sequences in monoscopic viewing conditions. In this experiment, the objective and subjective methods are now evaluated with the temporal domain. In these conditions, the objective is to determine:

- whether ACR-HR is appropriate for the assessment of different DIBR;
- the required number of participants for such a subjective assessment test;
- whether the results of the subjective assessments are consistent with the objective evaluations.

The following subsections describe the results of Experiment 2. The first part addresses the results of the subjective assessments and the second part presents the results of the objective evaluations.

6.6.1 Subjective tests

In the case of video sequences, only the ACR-HR test was conducted, as explained in Section 6.4. Table 6.11 shows the algorithms' ranking from the obtained subjective scores. The ranking order differs from the one obtained with ACR-HR test in the still image case.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
ACR-HR	3.523	3.237	2.966	2.865	2.789	2.956	2.104
Rank order	1	2	3	5	6	4	7

Table 6.11: *Rankings of algorithms according to subjective scores (video sequences)*

Although the values allow the ranking of the algorithms, they do not directly provide knowledge on the statistical equivalence of the results. Table 6.12 depicts the results of the Student's t-test processed with the values. Compared to the ACR-HR test with still images detailed in Table 6.6, distinctions between algorithms seem to be more obvious. The statistical significance of the difference between the algorithms, based on the ACR-HR scores, exists and seems clearer for video sequences than for still images. This can be explained by the exhibition time of the video sequences: watching the whole video, observers can refine their judgment, contrary to still images. Note that the same algorithms were not statistically differentiated: A4, A3, A5 and A6. As a conclusion, though more

	A1	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*
A1		↑(7)	↑(3)	↑(3)	**2**	↑(3)	↑(1)
A2	↓(7)		↑(2)	↑(2)	↑(1)	↑(2)	↑(1)
A3	↓(3)	↓(2)		**o(>32)**	↑(9)	**o(>32)**	↑(1)
A4	↓(3)	↓(2)	**o(>32)**		**o(>32)**	**o(>32)**	↑(1)
A5	↓(2)	↓(1)	↓(9)	**o(>32)**		↑(15)	↑(1)
A6	↓(3)	↓(2)	**o(>32)**	**o(>32)**	↑(15)		↑(1)
A7	↓(1)	↓(1)	↓(1)	↓ (1)	↓(1)	↓(1)	

Table 6.12: *Results of Student's t-test with ACR-HR results (video sequences). Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.*

than 32 participants are required to perform all the distinctions in the tested conditions, the ACR-HR test with video sequences gives clearer statistical differences between the algorithms than the ACR-HR test with still images. This suggests that new elements allow the observers to make a decision: existence of flickering, exhibition time, etc.

6.6.2 Objective measurements

The results of this subsection concern the measurements conducted over the entire synthesized sequences. The objective is to determine the consistency between the subjective assessments and the objective evaluations, and the most consistent objective metric, in the context of video sequences.

As in the case of still images studied in the previous section, the rankings given by the objective metrics in Table 6.13 are consistent with each other. Besides, the correlation coefficients between objective metrics are very close to the figures depicted in Table 6.10, and so they are not presented here. As with still images, the difference between the subjective test based ranking and the ranking from the objective scores is noticeable. Again, the algorithm given as the worst (A1) by the objective measurements, is the one observers preferred. This can be explained by the fact that A1 performs the synthesis on a cropped image, and then enlarges it to reach the original size. Consequently, signal-based metrics penalize it though it gives good perceptual results.

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.523	3.237	2.966	2.865	2.789	2.956	2.104
Rank order	1	2	3	5	6	4	7
PSNR	19.02	24.99	23.227	25.994	26.035	26.04	20.89
Rank order	7	4	5	3	2	1	6
SSIM	0.648	0.844	0.786	0.859	0.859	0.859	0.824
Rank order	7	4	6	1	1	1	5
MSSIM	0.664	0.932	0.825	0.948	0.948	0.948	0.888
Rank order	7	4	6	1	1	1	5
VSNR	13.14	20.41	18.75	21.786	21.965	21.968	20.73
Rank order	7	5	6	3	2	1	4
VIF	0.129	0.393	0.313	0.423	0.423	0.424	0.396
Rank order	7	5	6	2	2	1	4
VIFP	0.153	0.415	0.342	0.446	0.446	0.446	0.419
Rank order	7	5	6	1	1	1	4
UQI	0.359	0.664	0.58	0.598	0.598	0.598	0.667
Rank order	7	5	6	3	3	3	1
IFC	0.779	2.399	1.926	2.562	2.562	2.564	2.404
Rank order	7	5	6	2	2	1	4
NQM	8.66	15.933	13.415	16.635	16.739	16.739	10.63
Rank order	7	4	5	3	1	1	6
WSNR	14.41	20.85	18.853	21.76	21.839	21.844	16.46
Rank order	7	4	5	3	2	1	6
PSNR HSVM	13.99	19.37	18.361	21.278	21.318	21.326	16.23
Rank order	7	4	5	3	2	1	6
PSNR HSV	13.74	19.52	17.958	20.795	20.823	20.833	15.91
Rank order	7	4	5	3	2	1	6
VSSIM	0.662	0.879	0.809	0.899	0.898	0.893	0.854
Rank order	7	4	6	1	2	3	5
VQM	0.888	0.623	0.581	0.572	0.556	0.557	0.652
Rank order	7	5	4	3	1	2	6

Table 6.13: *Rankings according to measurements (video sequences)*

Table 6.14 presents the correlation coefficients between objective scores and subjective scores, based on the whole set of measured points.

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**	**UQI**
CC_{DMOS}	33.80	40.84	49.95	38.87	27.01	20.00	28.70
	IFC	**NQM**	**WSNR**	$PSNR_{HVSM}$	$PSNR_{HVS}$	**VSSIM**	**VQM**
CC_{DMOS}	21.68	41.15	27.98	31.47	29.76	39.37	43.41

Table 6.14: *Pearson correlation coefficients between DMOS and objective scores in percentage (video sequences).*

6.6.3 Conclusion

To conclude, when assessing views synthesized from DIBR in monoscopic viewing conditions, performances of objective metrics are different for video sequences and for of still images, in terms of correlation with subjective scores. Correlation coefficients between objective and subjective scores were higher in the case of video sequences, when comparing Table 6.14 with Table 6.8. However, human opinion also differed in the case of video sequences. For video sequences, perception-oriented metrics were the most correlated to subjective scores (also in video conditions). However, in either context, the tested metrics hardly reached substantial correlation with human judgment.The next section investigate the performances of the methods in stereoscopic viewing context.

6.7 Experiment 3: still images in stereoscopic conditions

In this experiment, still synthesized images quality is evaluated in stereoscopic conditions, as described in Sec. 6.4.2. In these conditions, the objective is to determine:

- whether ACR-HR is appropriate for the assessment of contents synthesized from different DIBR algorithms;
- the required number of participants for such a subjective assessment test;
- whether the results of the subjective assessments are consistent with the objective evaluations;
- whether the results are similar to the monoscopic conditions results.

The following subsections describe the results of Experiment 3. The first part addresses the results of the subjective assessments and the second part presents the results of the objective evaluations.

6.7.1 Subjective tests

The seven DIBR algorithms are ranked according to the obtained ACR-HR, as depicted in Table 6.15. The first line gives the DMOS scores obtained through the MOS scores. The second line gives the ranking of the algorithms, obtained through the first line.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
ACR-HR	3.647	3.637	3.660	3.678	3.658	3.662	3.548
Rank order	5	6	3	1	4	2	7

Table 6.15: *Rankings of algorithms according to subjective scores (stereoscopic still images).*

The first comment regarding the results in Table 6.15 refers to the fact that the rankings of the algorithm, according to the subjective scores, are completely different from the rankings obtained in monoscopic conditions, in Table 6.5 from Experiment 1. In particular, A1 was always ranked as the best algorithm in monoscopic conditions. It is ranked as 5^{th} in stereoscopic conditions. This can be explained by the discomfort produced by the proposed stereopairs. Indeed the stereopairs presented to the observers were made up of one original acquired view and its stereo-compliant synthesized view. Considering the interpolation strategy used in A1 (the borders of the image are cropped and then an interpolation allows to reach the original size of the image), we can observe that objects are shifted. This shift is assumed to be the cause of discomfort(Sec. 2.1.3), since corresponding objects will have too large disparity values. On the other hand, algorithms that were not ranked as the best in monoscopic conditions are better ranked in stereoscopic conditions. For instance, A6 or A3 that were ranked 4^{th} and 6^{th} respectively by the subjective scores in monoscopic conditions, are ranked 2^{nd} and 3^{rd} in stereoscopic conditions. The assumption is that the artifacts generated by these algorithms are more easily masked through human vision in stereoscopic conditions. They do not induce difficult-to-deal-with artifacts (in stereovision) such as shifts of objects.

Table 6.16 give the results of the Student's test from the ACR-HR scores for stereoscopic still images. The statistical significance of the differences between the algorithms are quite

different in the stereoscopic case than in the monoscopic case. In particular, the signs of the statistical differences are often opposite to those obtained in the monoscopic case, except for the case of $A7$. Considering the rankings of algorithms, observed above, this was expected. However, algorithms whose statistical significance was equivalent remain the same: this is observed for $A4$, $A5$ and $A6$. This suggests that these algorithms induced artifacts that were equivalent in both stereoscopic and monoscopic viewing conditions. Concerning the number of participants that allow a clear distinction between the algorithms, in most of the cases, less than 24 observers give clear distinctions. This is very different from the results obtained in Experiment 1, in monoscopic viewing conditions. Our assumption is that artifacts are differently perceived in stereoscopic viewing and in monsocopic viewing conditions. Artifacts that are not annoying in monoscopic viewing conditions, may be disturbing in stereoscopic conditions which explains the clear distinctions.

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		**o(>25)**	**o(>25)**	↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A2$	**o(>25)**		↓(<24)	↓ (<24)	↓(<24)	↓(<24)	↑(<24)
$A3$	**o(>25)**	↑(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↑(<24)	↑(<24)	↑(<24)		**o(>25)**	**o(>25)**	↑(<24)
$A5$	↑(<24)	↑(<24)	↑(14)	**o(>25)**		**o(>25)**	↑(<24)
$A6$	↑(<24)	↑(<24)	↑(18)	**o(>25)**	**o(>25)**		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓ (<24)	↓(<24)	↓(<24)	

Table 6.16: *Results of Student's t-test with ACR-HR results (stereo still images).*

6.7.2 Objective measurements

The results of this subsection concern the measurements conducted over the same selected "key" frames as those in Section 6.5. The objective is to determine the consistency between the subjective assessments and the objective evaluations, and the most consistent objective metric.

The first step consists in comparing the objective scores with the subjective scores (previously presented). The consistency between objective and subjective measures is evaluated by calculating the PLCC, as described in Eq. 6.2 for the whole fitted measured points (Eq. 6.1).

The coefficients are depicted in Table 6.17. Compared to Experiment 1 in monoscopic viewing conditions, the metrics are still not highly correlated to human judgment. Fig. 6.6 illustrates the differences between the PLCC obtained in monoscopic conditions and those obtained in stereoscopic conditions and confirms the previous comment. Depending on the objective metric, we observe that the PLCC is slightly improved.

Table 6.18 depicts the rankings of the algorithms obtained from the objective scores. The first line recalls the previously commented subjective results. In stereoscopic conditions, we observe that the rankings from the objective metrics are slightly closer to human judgment than in monoscopic conditions.

6.7.3 Conclusion

To conclude, the experiments in stereoscopic viewing conditions highlighted different observations:

- non-annoying artifacts in monoscopic conditions may not be assessed with the same quality especially in the case of shifting artifacts.
- correlation of objective metrics with subjective scores is not proved. The results suggest that the tested objective metrics are not sufficient to predict the quality of stereoscopic images.

The next section proposes a tool addressing the assessment of synthesized views.

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR HVSM	PSNR HVS
CC$_{DMOS}$	46.98	45.06	60.86	26.44	38.46	42.96	31.72	40.96	52.66	51.58	46.59	46.13

Table 6.17: *Pearson correlation coefficients between DMOS and objective scores in percentage (stereoscopic still images).*

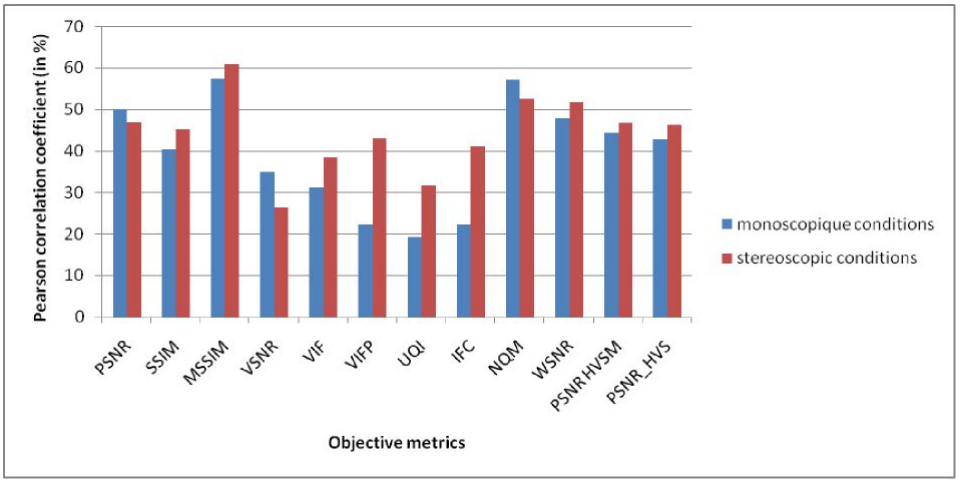


Figure 6.6: *Comparison of Pearson linear correlation coefficients in monoscopic and stereoscopic conditions.*

	A1	A2	A3	A4	A5	A6	A7
DMOS	3.647	3.637	3.660	3.678	3.658	3.662	3.548
Rank order	5	6	3	1	4	2	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Rank order	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Rank order	7	4	6	1	2	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Rank order	7	4	6	1	2	3	5
VSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Rank order	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Rank order	7	5	6	2	3	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Rank order	7	5	6	2	3	1	4
UQI	0.352	0.672	0.589	0.606	0.605	0.606	0.673
Rank order	7	2	6	3	5	4	1
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Rank order	7	5	6	2	3	1	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Rank order	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Rank order	7	4	5	3	2	1	6
PSNR HSVM	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Rank order	7	4	5	3	2	1	6
PSNR HSV	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Rank order	7	4	5	3	2	1	6

Table 6.18: *Rankings according to measurements (stereoscopic still images).*

Figure 6.7: *Synthesized views - Frame 45 - view 10 - Book Arrival.*

6.8 Our proposal: an edge-based structural distortion indicator

Most of the proposed metrics are inspired from 2D commonly used quality metrics. Yet, the latter were originally designed to address 2D compression distortions which are different from the distortions related to DIBR processes, as depicted in Fig. 6.7 and Fig. 6.8. Fig. 6.7 and Fig. 6.8 give examples of distortions. The synthesized views depicted in these figures were obtained through different DIBR algorithms. As it can be observed, the distortions are located around the edges of the arms for Fig. 6.7 and around the edges of the face for Fig. 6.8. This corresponds to strong depth discontinuities. Thus, artifacts related to DIBR systems are mostly located in specific areas that are the disoccluded regions. They are not scattered in the entire image such as specific 2-D video compression distortions.

2D commonly used quality metrics indicate the presence of errors but not necessarily the degree of visual annoyance. The previous study showed that they are not sufficient to assess the visual quality of synthesized views. We propose an edge-based method that indicates the level of structural degradation in the synthesized image.

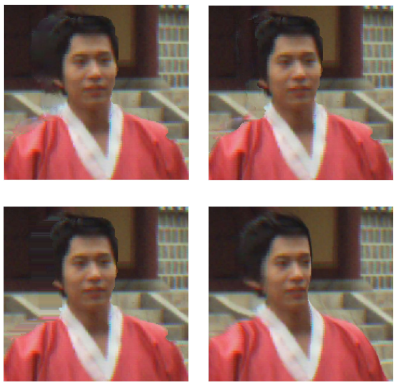


Figure 6.8: *Synthesized views - Frame 112 - view 6 - Lovebird1.*

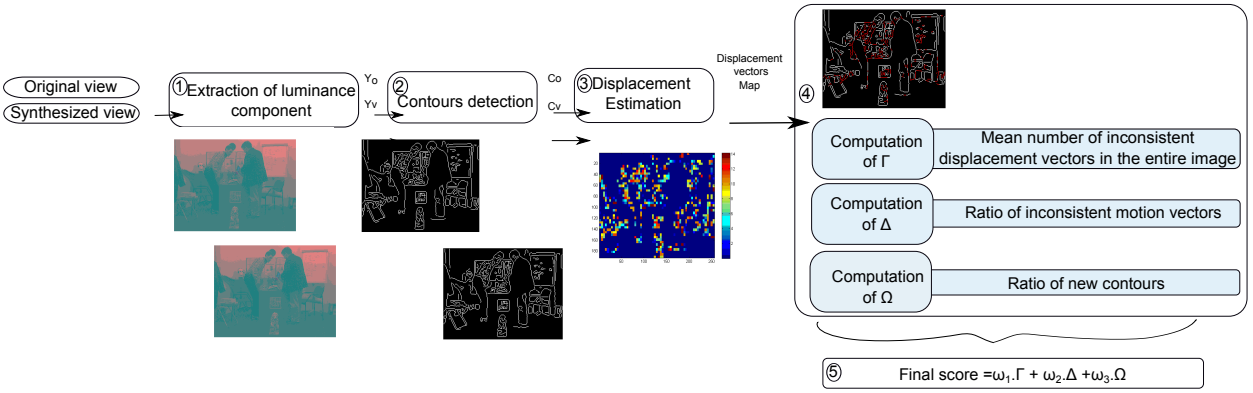


Figure 6.9: *Overview of the proposed method.*

Our proposed method is based on the analysis of the synthesized view edges compared to the edges of the original acquired view.

6.8.1 Proposed indicator

Our proposed method is based on the observation of the nature and of the location of the artifacts, as pointed out in the previous study. The method relies on the analysis of the synthesized view edges in comparison to the original image edges. More than a quality measure, it is an indicator of structural distortion. Let I_o be the original view, I_v is the virtual synthesized view. As depicted in Fig. 6.9, first step consists in extracting the luminance component of I_o and I_v, referred as Y_o and Y_v respectively. At the second step a Canny edge detector is applied on Y_o and Y_v. Let $\mathscr{C}_o$ and $\mathscr{C}_v$ be the resulting extracted contours. At the third step a displacement vector estimation is performed between $\mathscr{C}_o$ and $\mathscr{C}_v$. The resulting displacement vectors map is processed at the fourth step. Three parameters are computed: the mean ratio of inconsistent displacement vectors per contour pixel that is denoted as Γ, the ratio of pixels in the contour $\mathscr{C}_v$ having at least one inconsistent displacement vector that is denoted as Δ, the ratio of new pixels in the

contour $\mathscr{C}_v$ that is denoted as Ω. They are defined as follows:

$$\Gamma = \frac{1}{|\mathscr{C}_v|} \sum_{c=1}^{|\mathscr{C}_v|} \gamma_c \tag{6.3}$$

where γ_c is the ratio of inconsistent displacement vectors for the pixel $\mathscr{C}_v(c)$. It is defined as:

$$\gamma_c = \frac{1}{K} \sum_{n}^{N} \delta(i,n) \tag{6.4}$$

with N the size of the slide window, i and n are such as $Y_s(i) \in \mathscr{C}_v$ and $Y_v(n) \in \mathscr{C}_v$, K is a normalizing factor, and

$$\delta(i,n) = \begin{cases} 1, & \text{if } \widehat{\overrightarrow{M_i}\overrightarrow{M_n}} > Th \\ 0, & \text{otherwise} \end{cases} \tag{6.5}$$

M_i and M_n are the displacement vectors of $Y_v(i)$ and $Y_s(n)$, $\widehat{\overrightarrow{M_i}\overrightarrow{M_n}}$ is the angle formed between M_i and M_n, and Th is a threshold.
Let $\mathscr{NDV}$ be the number of pixels having at least one inconsistent displacement vector. This is expressed as:

$$\mathscr{NDV} = \sum_{c=1}^{|\mathscr{C}_v|} \phi(c) \tag{6.6}$$

with

$$\phi(c) = \begin{cases} 1, & \text{if } \gamma_c \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.7}$$

The ratio of pixels in the contour $\mathscr{C}_v$ having at least one inconsistent displacement vector, denoted as Δ is then expressed by:

$$\Delta = \frac{\mathscr{NDV}}{|\mathscr{C}_v|} \tag{6.8}$$

Ω, the ratio of new pixels in the contour $\mathscr{C}_v$ is expressed by:

$$\Omega = \frac{|\mathscr{C}_v| - |\mathscr{C}_o|}{|\mathscr{C}_v| + |\mathscr{C}_o|} \tag{6.9}$$

The final score is a weighting sum of the three parameters:

$$Indicator = 1 - (\alpha_1\Gamma + \alpha_2\Delta + \alpha_3\Omega) \tag{6.10}$$

In the experiments we used the combination$\{Th = 45°,\ \alpha_1 = 0.25,\ \alpha_2 = 0.25,\ \alpha_3 = 0.5\}$. The closer to 1 the final score, the less distorted the contours of the image.

6.8.2 Experimental results and discussion

The quality of the set of synthesized views used the previous section (Sec. 6.4.1) is assessed through commonly used metrics and through the proposed indicator. The obtained scores are fitted and scaled into a common MOS scale. Fig. 6.10, 6.11, 6.12 and 6.13 depict the quality scores of four synthesized views containing different type of distortions (color leak, blurry regions, and geometric distortions). In each figure, a bar plot gives the Mean Opinion Score (MOS), the PSNR fitted score, the fitted score the closest to the MOS

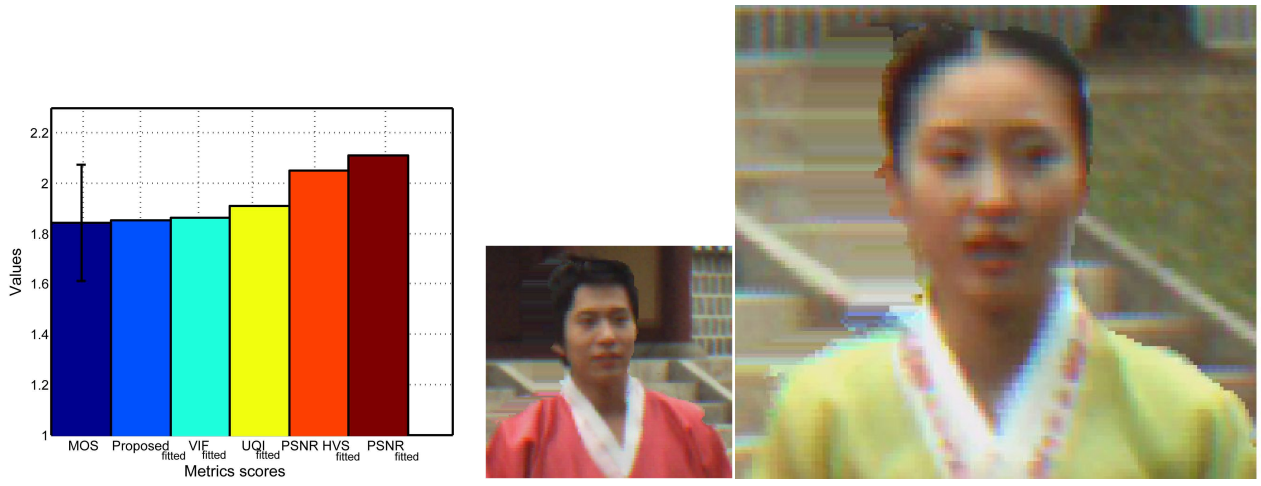


Figure 6.10: *Quality evaluation of a synthesized view (Frame 112 - view 6 - Lovebird1 rendered with [MSD+08])*

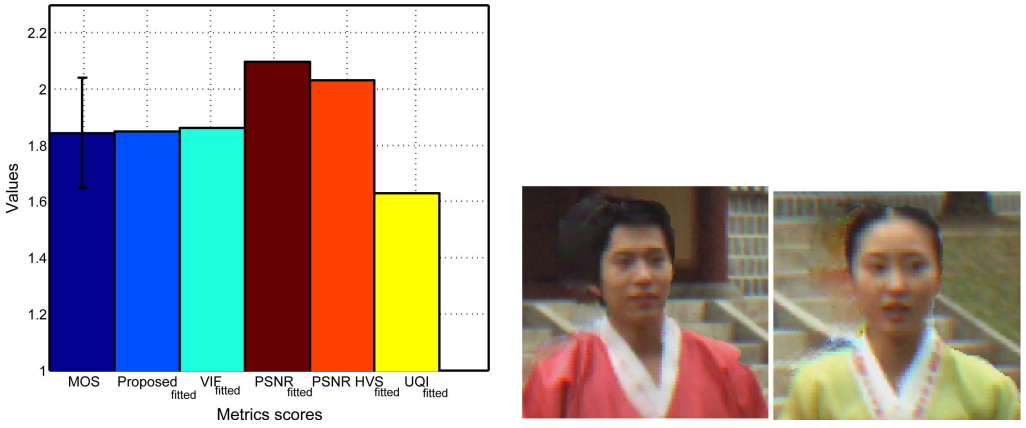


Figure 6.11: *Quality evaluation of a synthesized view (Frame 112 - view 6 - Lovebird1 rendered with [TFS+08])*

score among the objective metrics fitted scores, and the proposed indicator fitted score. The objective metric whose fitted score was closest to the MOS was not the same for all the figure, but we provide it for both figure. For example, in Fig. 6.10 the objective metric whose fitted score is the closest to the MOS is VIF, but in Fig. 6.12 it is PSNR-HVS. Also, particular regions of the synthesized views are provided in each figure. In the presented cases, PSNR shows the highest gap with the MOS score: even if a distortion is not perceptible, it contributes to the decrease of the quality score because its perceptual impact on the quality is not considered. The objective score that is the closest to the MOS is also provided. Although the closest objective score is different depending on the figure, it can be observed that it is a Human Visual System (HVS) based metric in both of the presented cases (Visual Information Fidelity (VIF) [SB06], PSNR-HVS [EAP+06] and Universal Quality Index [WB02] (UQI)). Contrary to the tested objective metrics, in both of the presented cases, our proposed indicator was close to the MOS or to the closest.

However, the proposed method does not assess the image quality, but it is able to detect

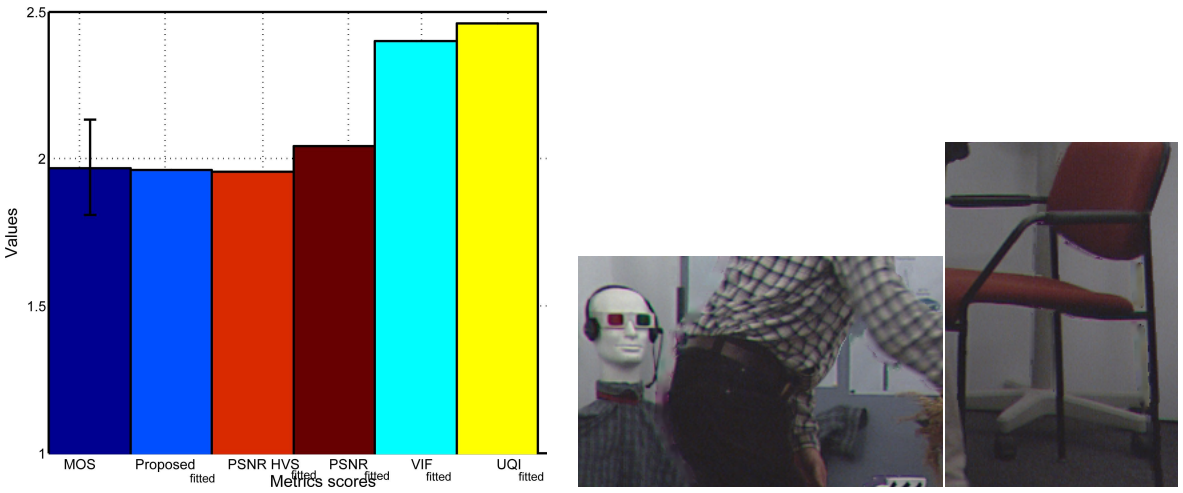


Figure 6.12: *Quality evaluation of a synthesized view (Frame 54 - view 10 - Book Arrival rendered with [TFS+08])*

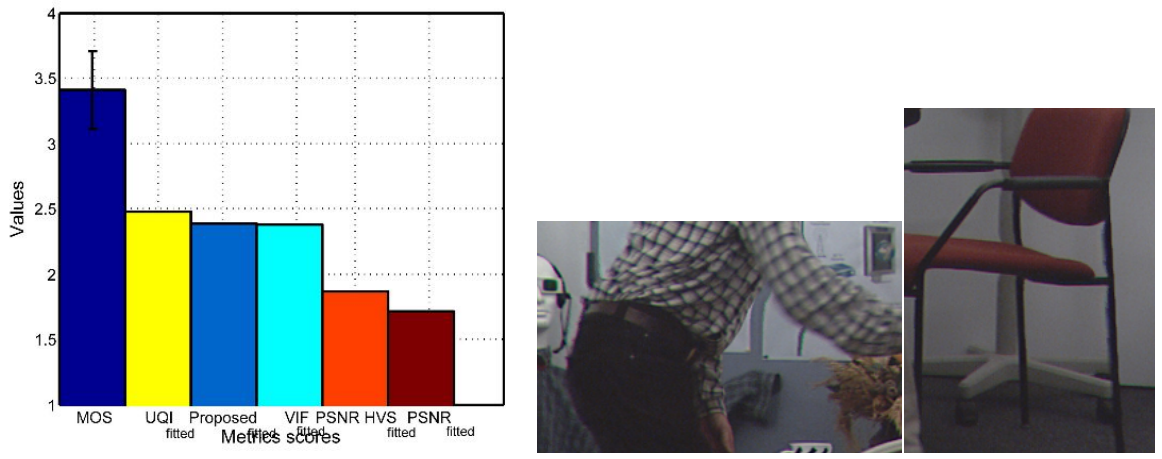


Figure 6.13: *Quality evaluation of a synthesized view (Frame 54 - view 9 - Book Arrival rendered with [Feh04])*

the structural distortions in this context. These results are very encouraging because at this stage the proposed method can only assess structural distortions. To be considered as a quality metric, the color consistency should also be analyzed and assessed. Moreover, the influence of weights α_i in Eq. (6.10) must be studied in future works.

6.9 Conclusion

The previous part showed that the synthesis process and the inpainting strategies induce specific distortions that are different from the commonly encountered artifacts in 2D imaging system. In this chapter, we thus questioned the relevance of the use of 2D usual quality assessment methods when addressing the quality of DIBR synthesized views.
We have presented our studies to investigate the reliability of usual methods of subjective and objective image/video quality assessment, in both monoscopic and stereoscopic viewing conditions. The results of these studies showed that non-annoying artifacts in monoscopic viewing conditions may be assessed with a worse quality when observed in stereoscopic viewing conditions, especially when corresponding objects in the stereopair lead to inconsistent and contradictory stereoscopic cues. These studies also showed that the required number of observers may be higher than the number recommended by VQEG (24).
The tested objective metrics obtained poor to medium correlation scores with human judgment in both monoscopic and stereoscopic viewing conditions.

Then, in this chapter, we also proposed a tool addressing the assessment of synthesized views, allowing the detection of DIBR-related distortions. This tool showed good abilities to predict human judgment in the presence of DIBR related artifacts. However, this is not a quality metrics yet, since several aspects such as color changes are not taken into account. This distortion indicator requires additional improvements that should be considered in future work.

The studies presented in this chapter helped understanding particular issues regarding the quality of reconstructed views in the context of use of DIBR. More precisely, understanding the origins of the distortions in synthesized views led us to devise more perceptually driven coding tools: the vocation of these coding tools is to deliver decompressed data that preserve the synthesized views visual quality. This is the subject of the next part of this thesis.

Part III

LAR-based MVD coding solutions

Various steps of the 3D Video processing chain can reduce the visual quality of the end user media, according to the previous parts of this thesis. These previous results showed that even if they are generated from original data, virtual viewpoints are impacted by different distortions depending on the inpainting strategy. In this part, we propose new tools for MVD coding whose aim is to enhance the visual quality of the synthesized views, based on the analysis of the sources of distortion at the synthesis process. The proposed tools rely on a 2D still image codec that is meant to preserve the essential depth information for a good quality of reconstructed virtual view.
Chapter 7 introduces the basics of this coding framework in the case of still image compression and investigates the performances of the method for depth maps compression. Then, Chapter 8 proposes a first depth-adapted coding solution that uses the associated decoded texture image for improving the prediction process. Finally, Chapter 9 proposes a second depth-adapted coding solution that enables scalable decoding thanks to its region-based prediction process.

CHAPTER 7

LAR codec

Before presenting our contributions to MVD coding, the method we based our solutions on has to be introduced. This is the goal of this chapter. The basics of Locally Adapted Resolution coding method are then presented in this chapter. Moreover, the second objective of this chapter is to explain the choice for this method as our basis. This is achieved through evaluations of LAR when encoding depth data. This chapter is organized as follows: the first section introduces the fundamentals of LAR method. The second section presents the evaluations of LAR method in the case of depth map coding.

7.1 Principles of LAR codec

The LAR compression method [DR99a, DR99b, DBBR07] was firstly designed for lossy gray scale images coding. It is based on the simple idea that an image can be considered as composed by a low resolution component and a component containing the details (see Fig. 7.1). For this reason, it is a two-layer codec : a spatial codec (also called flat codec) and a complementary spectral one, as depicted in Fig. 7.2. The spatial coder provides a low bit rate compressed image whereas the spectral one encodes the details. This basic scheme has been improved with numerous extensions. In this section, we first address the two coders (flat coder and spectral coder) before presenting the extension we will rely on in our contributions.

Figure 7.1: *Principle of LAR method.*

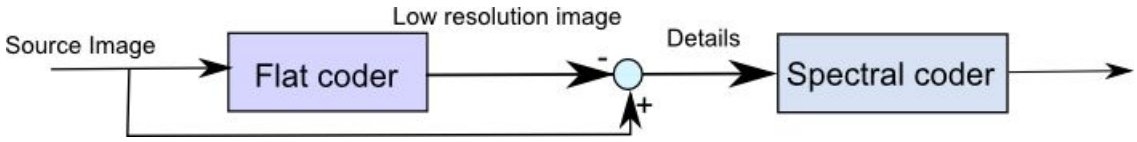


Figure 7.2: *General scheme of basic LAR codec.*

Original frame Variable block sizes representation

Figure 7.3: *Variable block sizes representation in LAR.*

7.1.1 Flat coder

The flat coder is based on the principle that local resolution should depend on local activity. It provides a flat representation of the original image: the image is segmented into blocks of various sizes (from 16×16 to 2×2) and each block is assigned the mean value of its pixels. The segmentation is driven by a quad-tree decomposition, dependent on a local gradient estimation. Consequently, small blocks of the representation are located on contours and large ones suit homogeneous areas, as illustrated on Figure 7.3. Perceptible blocks artifacts in homogeneous areas are easily removed by an efficient post-processing. This coder is dedicated to low bit-rate image coding. The following descriptions rely on the explanations provided in [Bab05]. An overview of the Flat coder is illustrated in Fig. 7.4.

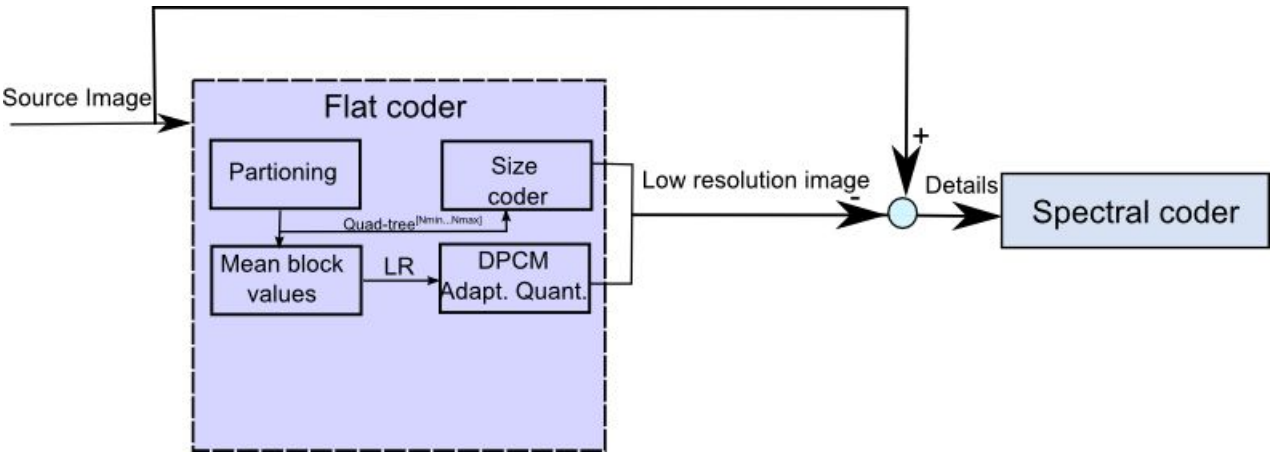


Figure 7.4: *General scheme of Flat codec layer.*

Quad-tree partioning

The quad-tree decomposition relies on a homogeneity criterion. Let $\textbf{Quad-tree}^{[N_{max}...N_{min}]}$ be the quad-tree partition, where N_{max} and N_{min} are the maximum and minimum allowed block sizes respectively. Let $I(x,y)$ be the pixel of coordinates (x,y) in image I and $I(b^N(i,j))$ is the block $b^N(i,j)$ of size $N \times N$ in image I, described as follows:

$$b^N(i,j) = \{(x,y) \in N \times N \mid N \times (i+1), \text{ and } N \times j \leq y \leq N \times (j+1)\} \tag{7.1}$$

The quad-tree decomposition is based on the detection of local activity. Considering a support, the difference between its maximal luminance value and its minimal luminance value is computed. For a given partition $\textbf{Quad-tree}^{[N_{max}...N_{min}]}$ of image I, for any pixel $I(x,y)$, the size of the block it belongs to in the partition is expressed as follows:

$$Size(x,y) = \left\{ \begin{array}{l} N \in [N_{max} \dots N_{min}[\quad \text{if} \\ \left| \begin{array}{c} |max\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) - min\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) \leq Y \\ \text{and if } \exists (k,m) \in \{0,1\}^2 \\ |max\big(I(b^N(\lfloor \frac{x+k}{N/2} \rfloor, \lfloor \frac{y+m}{N/2} \rfloor))\big) - min\big(I(b^N(\lfloor \frac{x}{N} \rfloor, \lfloor \frac{y}{N} \rfloor))\big) > Y \end{array} \right| \\ N_{min} \qquad \text{otherwise.} \end{array} \right. \tag{7.2}$$

where $min\big(I(b^N(i,j))\big)$ and $max\big(I(b^N(i,j))\big)$ are the minimal and the maximal values of block $I(b^N(i,j))$ respectively, and Y is the homogeneity threshold. The threshold value used to perform the quad-tree decomposition influences the final representation. It directly influences the sensitivity of the detection of homogeneous areas.

Averaging of blocks

The low resolution component of image I (i.e. the flat image) is obtained by averaging luminance values of each block of the quad-tree. Let LR be the low resolution image, each pixel $LR(x,y)$ is expressed as:

$$LR(x,y) = \frac{l}{N^2} \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} I(\lfloor \frac{x}{N} \rfloor \times N + k, \lfloor \frac{y}{N} \rfloor \times N + m) \tag{7.3}$$

where $N = Size(x,y)$.

7.1.2 Spectral coder

The encoding scheme used in the spectral coder relies on the implementation of a DCT adapted to the partition $\textbf{Quad-tree}^{[N_{max}...N_{min}]}$. Fig. 7.5 depicts the principles of this codec layer. The coefficient coding is computed through intra-block zigzag scanning, then non zero values are encoded through "run length" (RLC), including specific tags for maximal run length. Finally, the quantization matrix is adapted to block size, as explained in the next paragraph.

Quantization

In the basic LAR framework, the quad-tree partition $\textbf{Quad-tree}^{[N_{max}...N_{min}]}$ controls the quantization. That is to say that it is based on the assumption that large blocks require fine quantization (in uniform areas, human vision is highly sensitive to brightness variations) while coarse quantization (low sensitivity) is sufficient for small blocks. Block values are

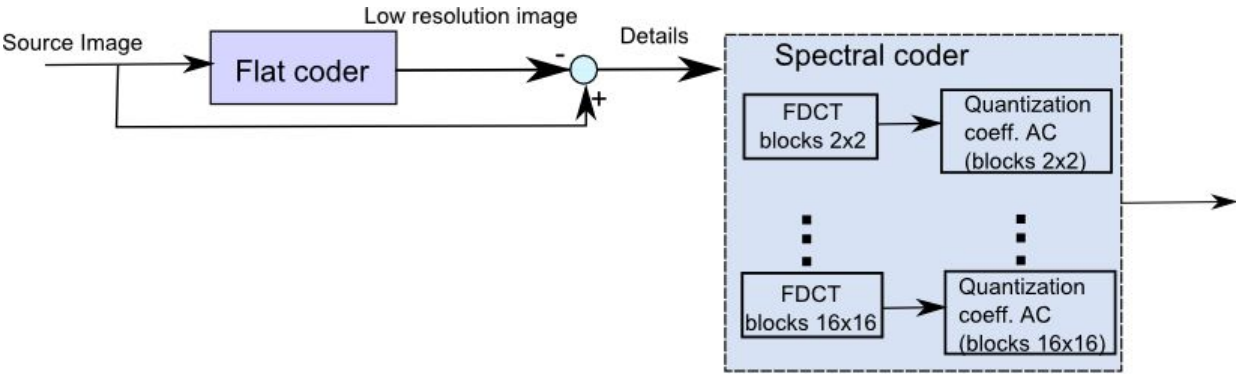


Figure 7.5: *General scheme of Spectral codec layer.*

encoded through a DPCM scheme. More precisely, the quantization step is calculated considering the fact that visual degradations related to a block are inversely proportional to the size of this block [BSB97]. For this reason, a relationship is defined between blocks of size $N \times N$ and blocks of size $N/2 \times N/2$: the quantization step q_N, for blocks of size $N \times N$ is defined as follows:

$$q_N = \frac{q_{N/2}}{2} \tag{7.4}$$

7.1.3 Pyramidal profile

The basic scheme of LAR method has led to many extensions. One of them is the pyramidal profile. Originally built to both increase scalability capacity and address lossless compression, multi-resolution extensions of the basic LAR called Interleaved S+P [BDR05, PBD+08] and RWHaT+P [DBBC08] were proposed. Since our contributions will rely on Interleaved S+P extension, in the following, the term "pyramidal profile" will refer to Interleaved S+P. To fit the Quadtree partition, dyadic decomposition is carried out. The first and second layers of the basic LAR are replaced by two successive pyramidal decomposition processes. However the image representation content is preserved. One pyramid is dedicated to the representation of the low resolution image and the second is dedicated to that of the details component. Fig. 7.6 shows an overview of the pyramidal decomposition. In the following, we first describe a typical pyramid construction. Then its representation using the S transform is presented. The typical pyramid reconstruction at the decoder side is then addressed.

Pyramid construction

The S+P transform (S transform + Prediction) is used in order to allow the decorrelation of the picture. It is briefly presented here because we assume it can help in the understanding of the coding artifacts, in our next studies. S+P transform is meant to allow lossless representation of images and scalable transmission of compressed data. The pyramid, built from image I, consists of a set of images, noted as $\{L_l\}_{l=0}^{l_{max}}$, as a multi-resolution representation of the image, where l_{max} is the top of the pyramid and $l = 0$ is the lowest level, i.e. the full resolution image. At each level, the image is expressed by:

$$\begin{cases} l = 0, & L_0(i,j) = I(i,j), \\ l > 0, & L_l(i,j) = \lfloor \frac{L_{l-1}(2i,2j) + L_{l-1}(2i+1,2j+1)}{2} \rfloor, \end{cases} \tag{7.5}$$

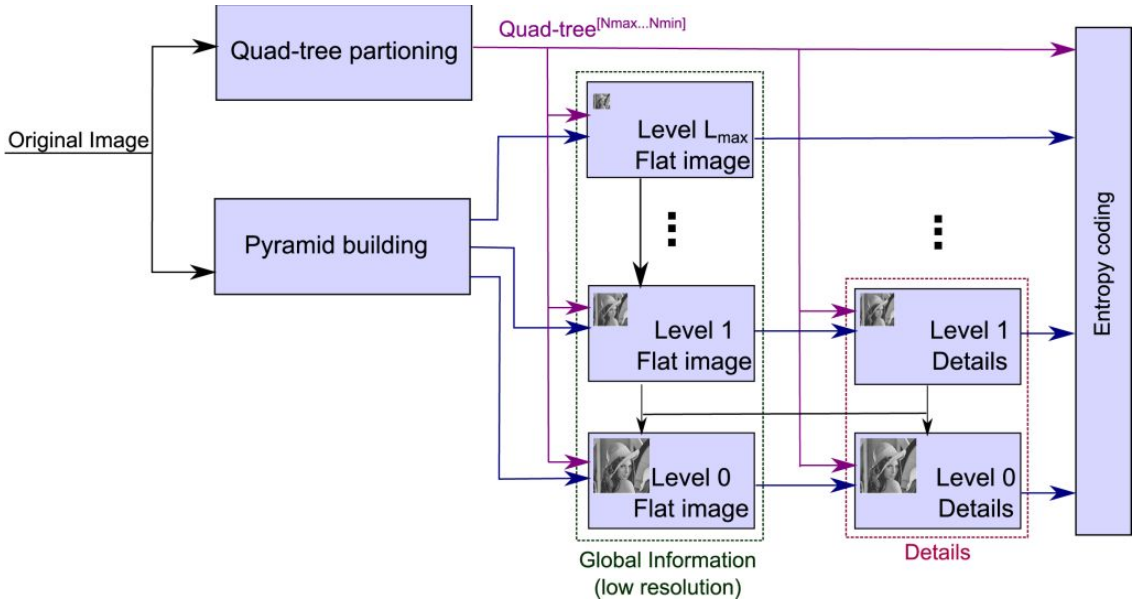


Figure 7.6: *LAR pyramidal decomposition.*

S-transformed pyramid

The S-transform is applied on the vectors formed by the two diagonally adjacent pixels in a 2x2 block as expressed in Eq. 7.6. The term "interleaved" of the method refers to the fact that the transformation of the second diagonal can be seen as a second S-pyramid.

$$\begin{cases} z_0 = \lfloor (u_0 + u_1)/2 \rfloor, \\ z_1 = (u_1 - u_0). \end{cases} \tag{7.6}$$

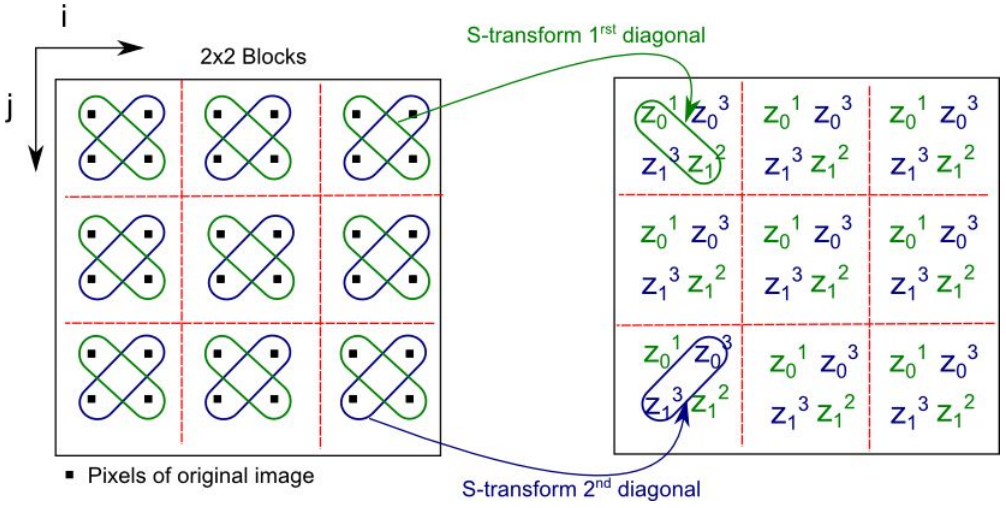


Figure 7.7: *S-transform scheme.*

Fig. 7.7 depicts an overview of the S-transform.There are three passes that are represented as exponent of the coefficients. First, the S-transform is applied on pixels $L_l(2i, 2j)$ and $L_l(2i+1, 2j+1)$ following the expression:

$$\begin{aligned} z_0^{l,1}(2i, 2j) &= \lfloor \tfrac{L_l(2i,2j)+L_l(2i+1,2j+1}{2} \rfloor = L_{l+1}(i, j), \\ z_1^{l,2}(2i+1, 2j+1) &= L_l(2i, 2j) - L_l(2i+1, 2j+1). \end{aligned} \tag{7.7}$$

Then, the transform is applied on the pixels from the second diagonal of blocks such as:

$$\begin{aligned} z_0^{l,3}(2i+1,2j) &= \lfloor \tfrac{L_l(2i,2j+1)+L_l(2i+1,2j}{2} \rfloor, \\ z_1^{l,3}(2i,2j+1) &= L_l(2i+1,2j) - L_l(2i,2j+1). \end{aligned} \tag{7.8}$$

The pyramid already stores the $z_0^{l,1}$ coefficient which is an average value of the diagonal of one 2×2 block. This value is obtained in the upper level of the pyramid, L_{l+1}. For this reason, only three types of coefficients have to be estimated for each level: $z_1^{l,2}$, $z_0^{l,3}$, $z_1^{l,3}$. The first layer coder (Flat layer) builds the first pass of the pyramid used by the Interleaved S+P. It decomposes each pixel of a given layer into a 2×2 block into a lower level according to the information given by the quad-tree. To perform a lossless compression, the second layer coding (Details) performs a second pass on this pyramid. It decomposes every pixels that have not been decomposed previously.

Pyramid reconstruction at the decoder side

From the top of the pyramid, the reconstruction of the lower levels only requires the gradient values, by prediction. On the highest level of the pyramid, the first pass is applied so that the $L_{l_{max}}(i,j) = z_0^{l_{max}-1,1}(2i,2j)$ coefficient values are predicted.
The flat image is reconstructed according to the following principle: for a given level of the pyramid, a block is processed (i. e. decomposed) only if the corresponding block size is lower or equal to the level size, that is to say when $Size(i \times l, j \times l) \leq 2^l$. If $Size(i \times l, j \times l) > 2^l$, the values of the block are copied according to:

$$L_l(2i,2j) = L_l(2i+1,2j) = L_l(2i,2j+1) = L_l(2i+1,2j+1) = L_{l+1}(i,j) \tag{7.9}$$

Recovering the low resolution image requires the prediction and decoding of S coefficients that have been actually transmitted to the decoder. Considering the decomposition of the first S-pyramid, as said before, $z_0^{l,1}$ is known from the upper level of the pyramid. Coefficient $z_1^{l,2}$ is predicted by $\tilde{z}_1^{l,2}$ according to [BDR05]:

$$\begin{aligned} \tilde{z}_1^{l,2}(2i+1,2j+1) = \; & 2.1\big[0.9L_{l+1}(i,j) + \tfrac{1}{6}\big(L_l(2i+1,2j-1) \\ & + L_l(2i-1,2j-1) + L_l(2i-1,2j+1)\big) \\ & - 0.05\big(L_l(2i,2j-2) + L_l(2i-2,2j)\big) \\ & - 0.15\big(L_{l+1}(i,j+1) + L_{l+1}(i+1,j)\big) - L_{l+1}(i,j)\big)\big] \end{aligned} \tag{7.10}$$

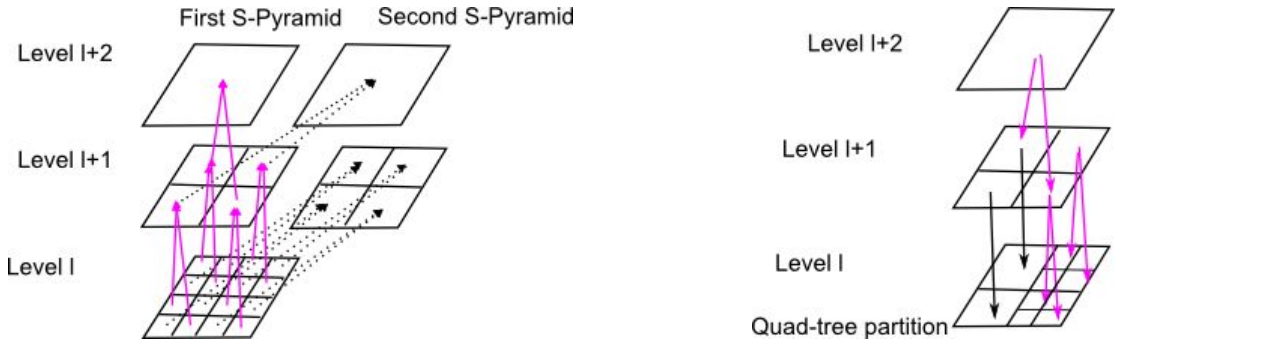
Considering the decomposition of the second S-pyramid, the Wu predictor[Wu97] is used. The predicted values of $z_0^{l,3}$, $z_1^{l,3}$ are respectively $\tilde{z}_0^{l,3}$, $\tilde{z}_1^{l,3}$:

$$\begin{aligned} \tilde{z}_0^{l,3}(2i+1,2j) = \; & \alpha_0 \tfrac{1}{4}\big(L_l(2i-1,2j+1) + L_l(2i,2j+2) \\ & + L_l(2i+2,2j) + L_l(2i+1,2j-1)\big) \\ & + \beta_0 \hat{z}_0^{l,1}(2i,2j), \end{aligned} \tag{7.11}$$

where $\hat{z}_0^{l,1}$ is the reconstructed value of $z_0^{l,1}$, $\alpha_0 = 0.25$ and $\beta_0 = 0.75$; and:

$$\begin{aligned} \tilde{z}_1^{l,3}(2i,2j+1) = \; & \alpha_1\big(L_l(2i-1,2j+1) + L_l(2i,2j+2) \\ & + L_l(2i+1,2j-1) + Yl(2i+2,2j)\big) \\ & - \beta_1\big(L_l(2i-1,2j) + L_l(2i-1,2j+2) \\ & - L_l(2i,2j-1) - \hat{L}_l(2i,2j+1). \end{aligned} \tag{7.12}$$

where $\hat{L}_l(2i,2j+1)$ is the Wu predictor [Wu97] for third pass applied to the pixel $L_l(2i,2j+1)$, $\alpha_1 = \frac{3}{8}$, and $\beta_1 = \frac{1}{8}$.



Construction of the pyramid Decomposition: extraction of LAR block image data

Figure 7.8: *Construction and decomposition of the pyramid.*

This presentation of the prediction step, as part of the reconstruction of the low resolution image, mainly aimed at highlighting the sources of distortions of the decoded image. Fig. 7.8 illustrates the construction and the decomposition of the pyramid.

This section has presented the basic profile of LAR codec and its pyramidal profile. The next section addresses the performances of this method when applied to depth map compression.

7.2 Depth coding with LAR codec

Considering LAR coding framework, we assumed that this method was appropriate for depth map coding. This was based on the observation that depth maps contains sharp edges and smooth areas that could be well represented through the LAR quad-tree decomposition. To validate our assumption, we first evaluated LAR codec performances in its original version, without any contributions.

The goal of this section is to present this evaluation of LAR codec performances when applied on depth maps. Thus, only depth maps are encoded, while texture images remain as original. As explained in Chapter 6, since accurate depth data is essential for a good rendering quality of synthesized views and since the latter are destined to be actually observed by the users, it is essential to assess the quality of the synthesized views. In this section, a first part addresses the experimental protocols, then a study on specific parameters of LAR codec is presented (threshold Y in Sec. 7.2.3 and the quantization steps in Sec. 7.2.4). Indeed, LAR parameters are numerous and we only focused on the most appropriate for depth map coding in this study.

7.2.1 Global Protocol

The scheme presented in Fig. 7.9 defines the protocol: given two viewpoints (left and right) are used, with color images and associated depth maps for each viewpoint. Only depth maps go through the compression process, because the experiments aim at evaluating coding artifacts caused by depth compression. Depth maps are encoded. Then they are decoded and they are used to perform the intermediate view synthesis, together with the color images that remained original. As explained before, the virtual view synthesis is necessary in the evaluation of depth compression, because accurate depth data is essential

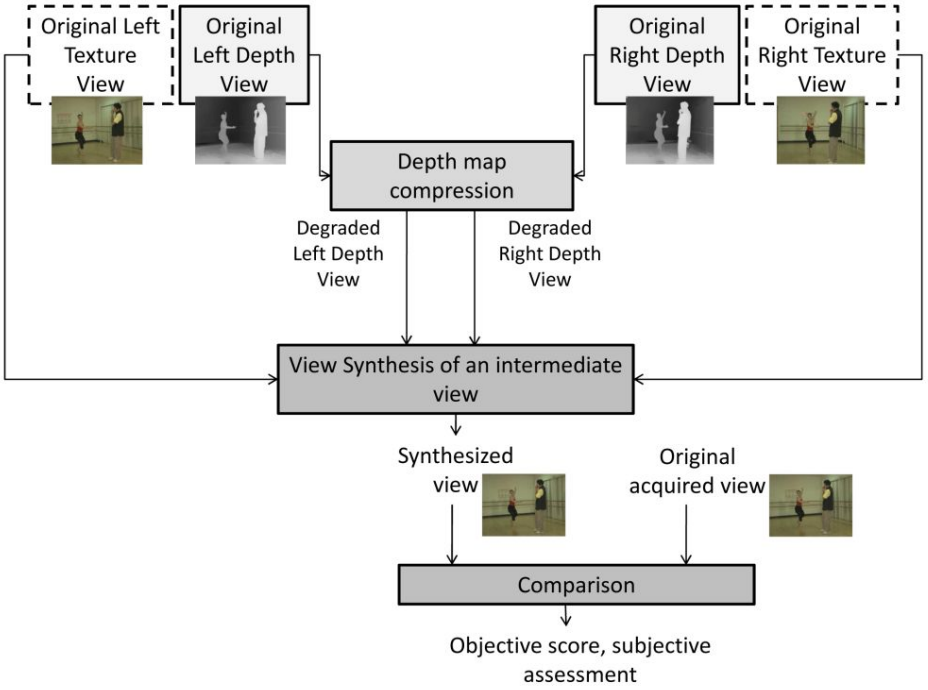


Figure 7.9: *Overview of the basic experimental protocol.*

for a good rendering quality of the synthesized view. View Synthesis Reference Software (VSRS, version 3.5) [TFS+08], provided by MPEG, is used for the view synthesis step; hence, decompressed depth maps are used for the virtual view synthesis. Three video sequences, namely *Breakdancers*, *Ballet*, *Book Arrival* are used in these experiments. In the case of *Breakdancers* and *Ballet* sequences, decoded depth maps of views 2 and 4 are used to synthesize view 3. In the case of *Book Arrival* sequence, decoded depth map of views 8 and 10 are used to synthesize view 9. Considering the results of the studies questioning the reliability of the objective metrics, PSNR score will be mentioned as an indicator of error rate for the evaluation of the reconstructed depth maps and synthesized views. Since this is not sufficient, visual observation is required: snapshots of reconstructed depth maps and synthesized views will be provided. Three different experiments were conducted in order to question:

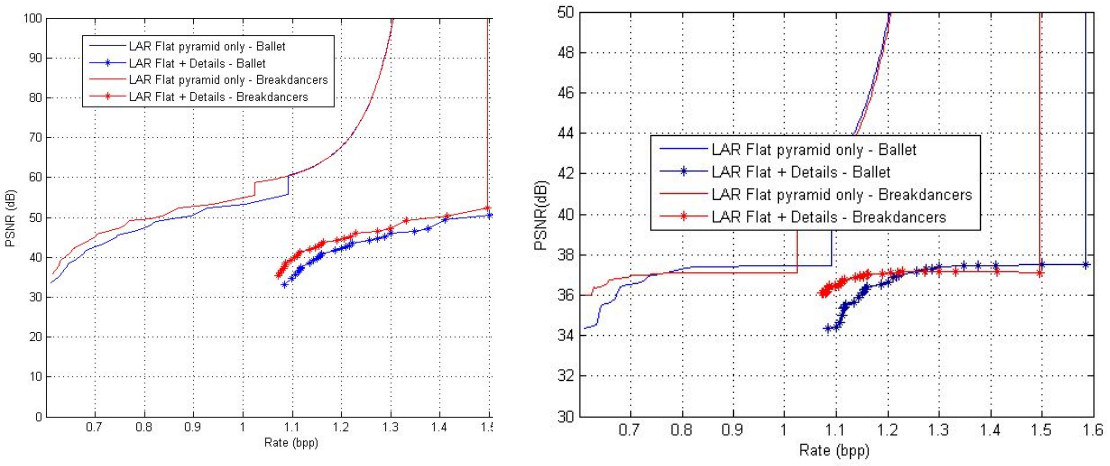
- the usefulness of the details pyramid for depth map coding in Sec. 7.2.2,
- the influence of threshold Y on the synthesized views quality in Sec. 7.2.3,
- the impact of the LAR quantization strategy on the synthesized views quality in Sec. 7.2.4.

These three experiments are discussed in the following subsections.

7.2.2 Flat and enhanced representations

This study questions the usefulness of the details pyramid for depth map coding. The assumption is that the essential depth information is already contained in the flat image, given the special features of depth maps (sharp edges and smooth areas). So, we compare

the coding performances when using only the flat coder and when using the flat coder and the texture coder together using the LAR pyramidal profile. In both cases, we use a quad-tree decomposition with threshold $Y = 1$. Quantization steps vary from 0 to 28. *Breakdancers*, *Ballet* are used in this experiment. Fig. 7.10 depicts the objective results.



(a) Distortion of depth maps. (b) Distortion of synthesized views with respect to original frame.

Figure 7.10: *Distortion according to the use or the details component.*

The curves showing the depth distortion (Fig. 7.10(a) as well as that of the synthesized distortion (Fig. 7.10(b)) lead to the same conclusion. The "*LAR Flat pyramid only*" curves are shifted on the left from the "*LAR Flat + Details*" curves, in both depth and synthesized view distortion cases. It appears that encoding the details pyramid is costly in terms of bit-rate but does not lead to a serious gain in terms of quality. $Y = 1$ is however a particular case, so tests also included other threshold values. Similar results were obtained for $Y = 11$, $Y = 21$ and $Y = 31$. Note that lossless compression can be reached with "*LAR Flat pyramid only*", $Y = 1$ and $QP = 0$. In this case, PSNR is infinite. These results prove that, based on the LAR conception of an image, the essential depth information is already contained in the low resolution component of the image. This was expected because depth maps represent smooth areas with sharp edges and few high frequency areas.

Fig. 7.11 depicts snapshots of the synthesized frames. In the particular case of $Y = 1$ and $QP = 0$, there was no difference between "*LAR Flat pyramid only*" and "*LAR Flat + Details*". However, when quantization is coarser, the synthesized views are more distorted in the case of "*LAR Flat + Details*" because the prediction errors are propagated in both pyramids. For this reason and also because of the bit-rate cost, we conclude that "*LAR Flat pyramid only*" is preferable to "*LAR Flat + Details*" for depth maps coding. In the following experiments "*LAR Flat pyramid only*" will be used as the reference.

Depth map bit rate can be controlled through two parameters that are the quantization step and threshold Y. Their impact on the synthesized views quality is studied in the next subsections.

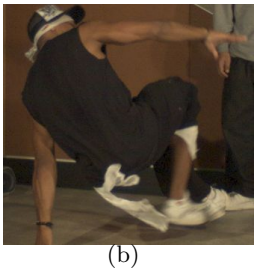
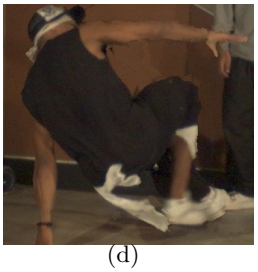
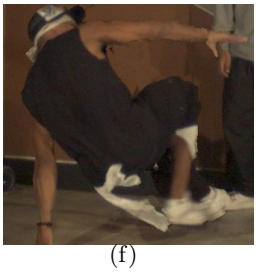
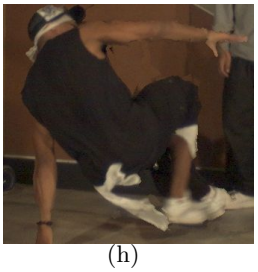
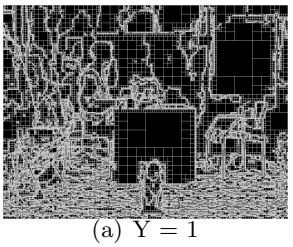
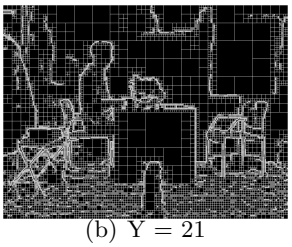
	Ballet	*Breakdancers*
Original	 (a)	 (b)
Synthesis from uncompressed data	 (c) PSNR: 37.42dB	 (d) PSNR: 37.08dB
Flat LAR	 (e) PSNR: 37.42dB 1.39bpp QP = 0 Y = 1	 (f) PSNR: 37.08dB 1.31bpp QP = 0 Y = 1
Flat LAR + Texture	 (g) PSNR: 37.42dB 1.92bpp QP = 0 Y = 1	 (h) PSNR: 37.08dB 1.81bpp QP = 0 Y = 1

Figure 7.11: *Synthesized frames.*



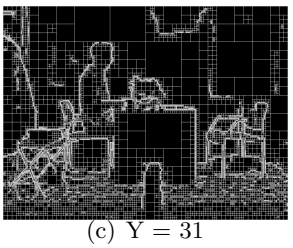
(a) Y = 1

Nb levels: 7
Blocks 2× 2: 47664
Blocks 4× 4: 10684
Blocks 8× 8: 2362
Blocks 16× 16: 457
Blocks 32× 32: 93
Blocks 64× 64: 7
Blocks 128× 128: 2



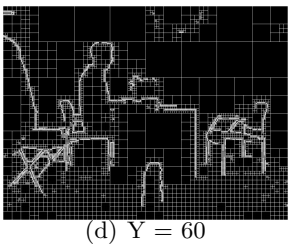
(b) Y = 21

Nb levels: 7
Blocks 2× 2: 12448
Blocks 4× 4: 6268
Blocks 8× 8: 3083
Blocks 16× 16: 607
Blocks 32× 32: 137
Blocks 64× 64: 27
Blocks 128× 128: 2



(c) Y = 31

Nb levels: 7
Blocks 2× 2: 10232
Blocks 4× 4: 4738
Blocks 8× 8: 2420
Blocks 16× 16: 635
Blocks 32× 32: 144
Blocks 64× 64: 34
Blocks 128× 128: 4



(d) Y = 60

Nb levels: 7
Blocs 2× 2: 5592
Blocs 4× 4: 2102
Blocs 8× 8: 1001
Blocs 16× 16: 811
Blocs 32× 32: 168
Blocs 64× 64: 38
Blocs 128× 128: 8

Figure 7.12: *Quad-tree decomposition for four different threshold values - Book Arrival.*

7.2.3 Threshold Y

The threshold value Y directly determines the quad-tree decomposition and thus the final image representation. Because the previous experiments confirmed that the use of "*LAR Flat pyramid only*" is more judicious than "*LAR Flat + Details*", the choice of the threshold is even more critical, considering the image representation. In this experiment, we use "*LAR Flat pyramid only*" to encode *Book Arrival* and *Breakdancers* depth maps, with $Y = 1$, $Y = 11$, $Y = 21$ and $Y = 31$. The quantization step varies from 0 to 28.
Fig. 7.12, 7.13 and 7.14 show the quad-trees obtained for different threshold values. It can be observed that for a large range of Y values (from 1 to 60 in this example), the main contours are preserved by the quad-tree representation. As illustrated by those figures, the lower the threshold value Y, the more sensitive the quad-tree is regarding the discontinuities of the depth maps: more discontinuities are detected since the number of small blocks is higher for low Y values. Consequently, a low threshold value leads to a large amount of small blocks, located around the boundaries.

Fig. 7.15 and 7.16 depict the rate-distortion curves for *Breakdancers* and *Book Arrival* sequences, respectively. For both sequences, we observe that the bit rate decreases when Y increases. This can be explained by the fact that high values of Y lead to more numerous large blocks and less erroneous small blocks. This is less costly because the total number of blocks decreases. Also, due to the averaging of blocks, the entropy decreases. Snapshots of depth maps (Fig. 7.17 and 7.19) show the influence of Y value on the depth map representation. Because of blocks averaging, depth values in smooth areas are modified. In other words, although the main contours of the scene objects are preserved by the quad-tree representation, for a large range of Y values, the depth structure of the scene is modified according to the block averaging. Since the pixels are thus assigned incorrect depth values, this will cause projection errors in the synthesis process. This assumption is confirmed in Fig. 7.18 and 7.20 that depict snapshots of the synthesized frames. In Fig. 7.18 it can be observed that the arms of the chair are exactly distorted according to the degradation of the corresponding depth maps of Fig. 7.17. One black piece of chair is slightly shifted in Fig. 7.18 with $Y = 60$, according the averaged depth block of corresponding pixels

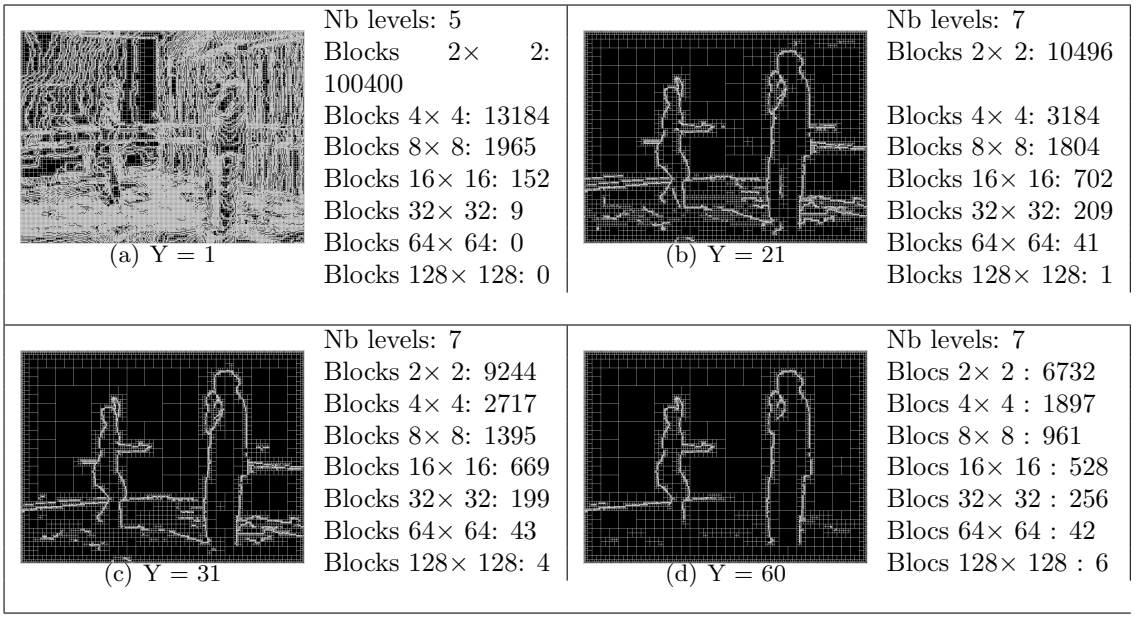


Figure 7.13: *Quad-tree decomposition for four different threshold values - Ballet.*

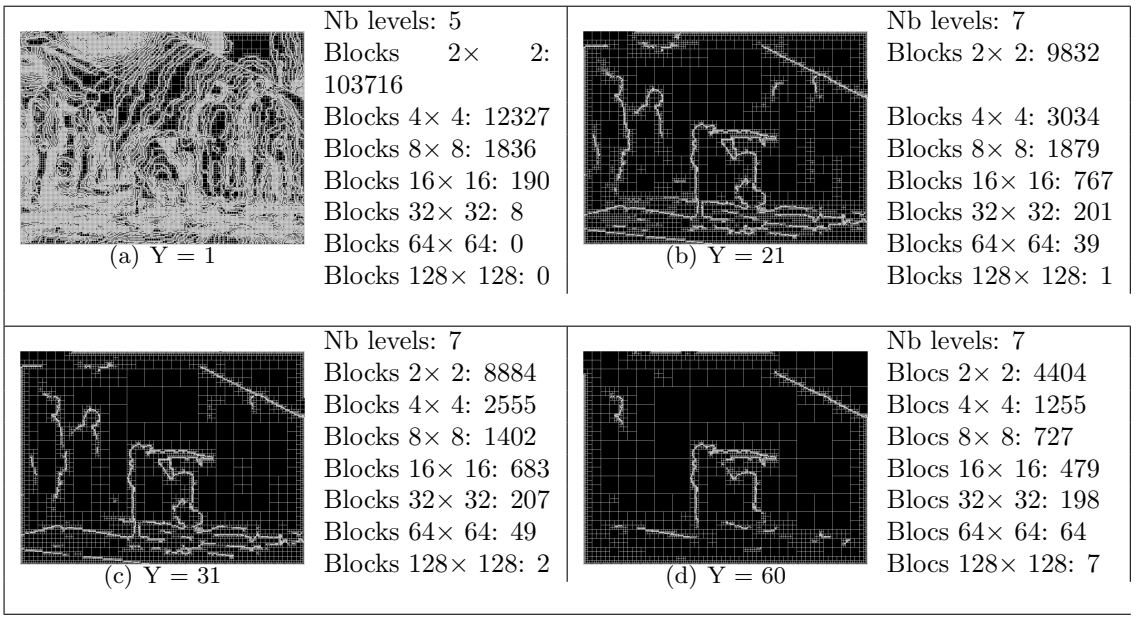
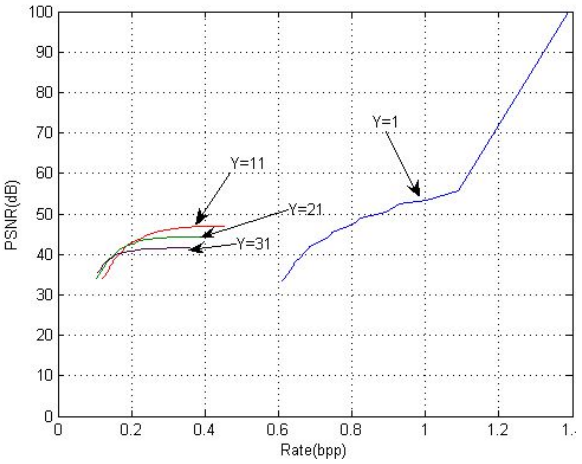


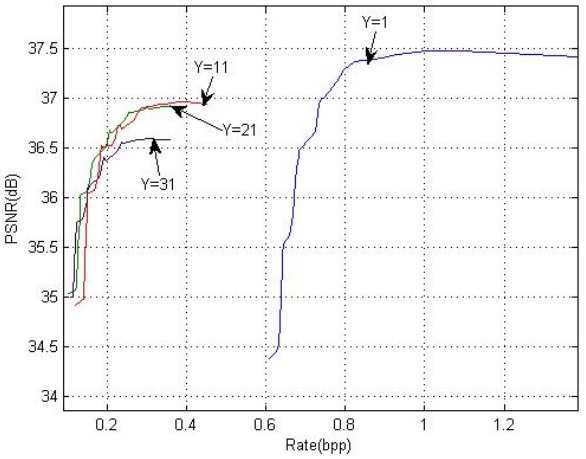
Figure 7.14: *Quad-tree decomposition for four different threshold values - Breakdancers.*

in 7.17 with $Y = 60$. Because of the block averaging, the corresponding depth pixels are wrong. Then, after the warping process, color pixels are wrongly projected, which explains the shifting. In Fig. 7.20, the distortions are perceptible around the feet of the dancer when Y increases. For the same reasons, color pixels are wrongly projected because of the incorrect depth values.

In conclusion, threshold Y determines the quad-tree decomposition and thus the representation of the scene depth. For a large range of Y values, the main contours of the image are preserved by the quad-tree decomposition. Moreover, although increasing threshold Y value can impact on the synthesized views quality, due to incorrect depth values in smooth areas, it allows bit rate savings.

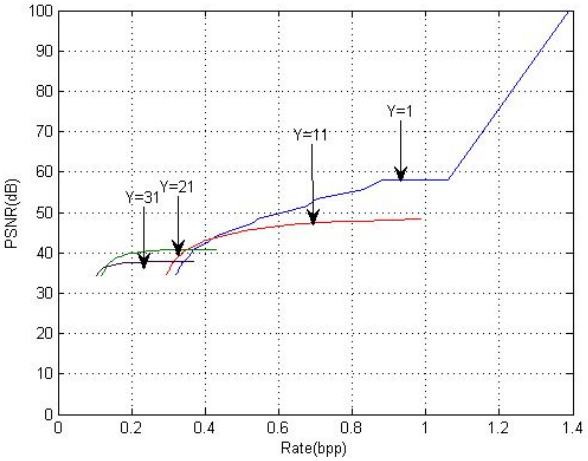


(a) Distortion of depth maps.

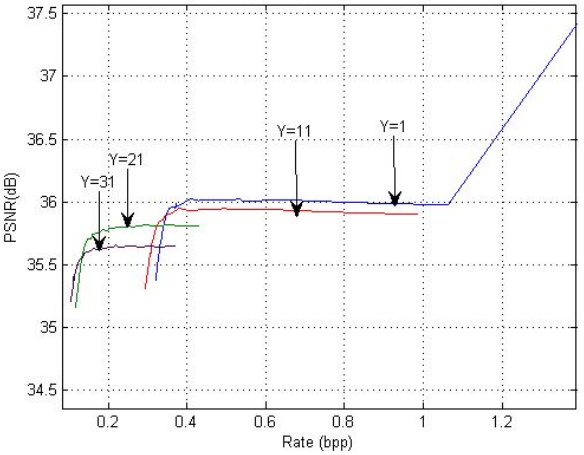


(b) Distortion of synthesized views with respect to original frame.

Figure 7.15: *Distortion depending on Y - Breakdancers.*

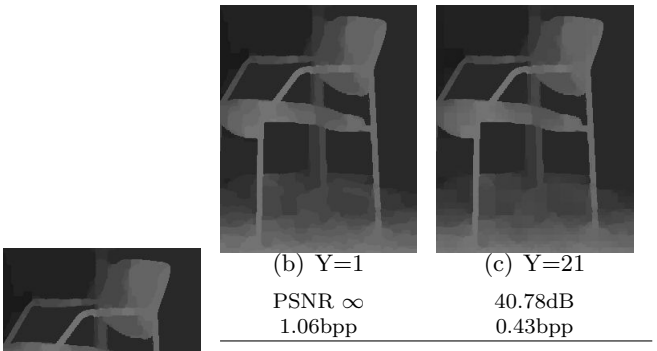


(a) Distortion of depth maps.



(b) Distortion of synthesized views with respect to original frame.

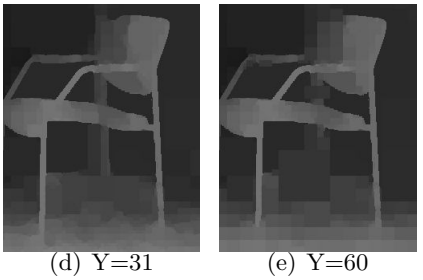
Figure 7.16: *Distortion depending on Y - Book Arrival.*



(a) Original (view 3, frame 0)

(b) Y=1
PSNR ∞
1.06bpp

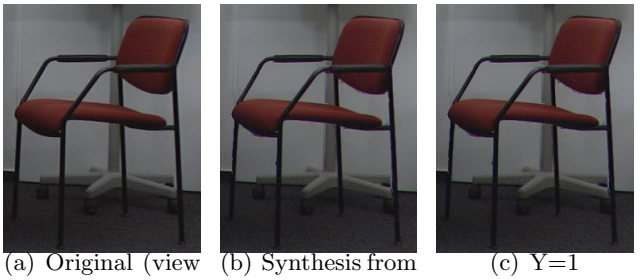
(c) Y=21
40.78dB
0.43bpp



(d) Y=31
37.77dB
0.37bpp

(e) Y=60
34.08dB
0.23bpp

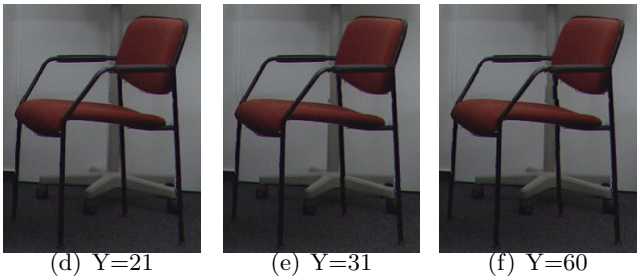
Figure 7.17: *Decoded depth maps - Book Arrival*



(a) Original (view 9, frame 0)

(b) Synthesis from original data
35.98dB

(c) Y=1
35.98dB
1.06bpp



(d) Y=21
35.81dB
0.43bpp

(e) Y=31
35.65dB
0.37bpp

(f) Y=60
35.28dB
0.23bpp

Figure 7.18: *Synthesized images - Book Arrival*

(a) Original (view 3, frame 0)

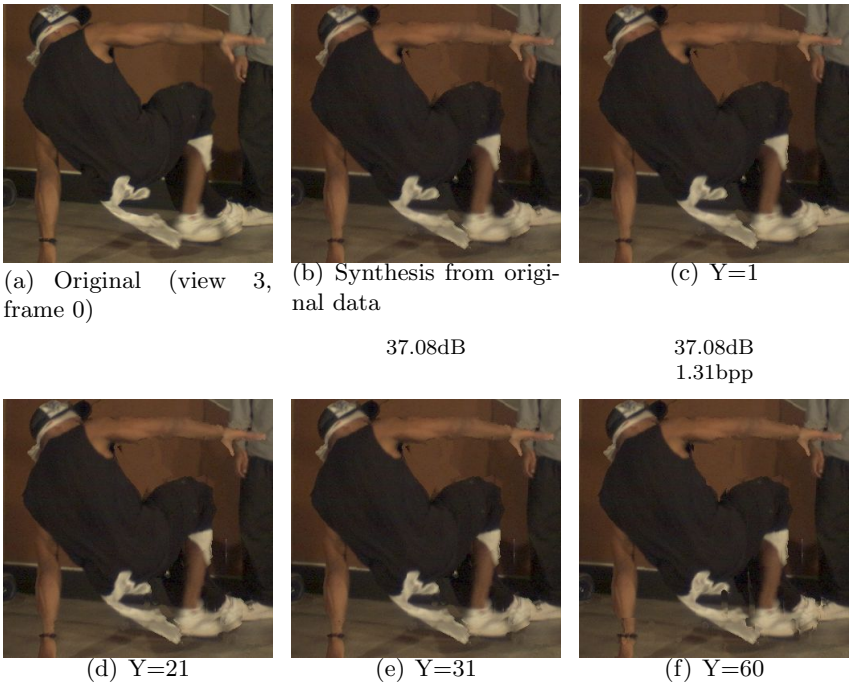
(b) Y=1
PSNR ∞
1.31bpp

(c) Y=21
44.69dB
0.33bpp

(d) Y=31
42.35dB
0.30bpp

(e) Y=60
35.83dB
0.19bpp

Figure 7.19: *Decoded depth maps - Breakdancers*



(a) Original (view 3, frame 0)

(b) Synthesis from original data
37.08dB

(c) Y=1
37.08dB
1.31bpp

(d) Y=21
36.47dB
0.33bpp

(e) Y=31
36.15dB
0.30bpp

(f) Y=60
35.64dB
0.19bpp

Figure 7.20: *Synthesized images - Breakdancers*

7.2.4 Quantization

In this experiment the effects of quantization on the rendered view quality are studied. The quantization parameter of LAR compression varies between 0 and 28, where 0 means the

highest quality and 28 corresponds to the lowest. The sequences used in this experiment are *Book Arrival* and *Breakdancers*. "*LAR Flat pyramid only*" is used to encode the depth maps, with $Y = 1$ and with $Y = 60$ and varying the quantization step from 0 to 28. Fig. 7.15 and 7.16 can be referred for the observation of PSNR when the quantization step varies. As expected, when the quantization step increases, the bit rate and PSNR scores decrease.

Figures 7.21 and 7.22 show a particular area of the decoded depth map for different threshold values ($Y = 1$ and $Y = 60$) and different quantization steps (noted QP) (5 and 25). The graphs show the depth values along the line number 395 of the image (arbitrarily chosen). It should be mentioned that for a quantization step equal to 0 and threshold $Y = 1$, the values of the decoded map fit exactly the original. The graphs show that the quantization adds a random noise around high discontinuities, no matter the value of the threshold Y. This is also noticeable on the snapshots of decoded depth maps. In particular, in Fig. 7.21, for $Y = 1$ and $QP = 25$ the contours of the chair appear noisy, because of wrong small block values. This leads to distortions in the synthesized view: the arms of the chair seem trimmed. The same occurrence can be observed in Fig. 7.22 for $Y = 1$ and $QP = 25$: in the synthesis, the dancer looks trimmed, or "crumbled" because of the wrong depth block values around its contour. The origin of this noise is assumed to be the result of prediction errors, from Eq. 7.10, Eq. 7.11 and Eq. 7.12. From top to bottom, the prediction of S coefficients may induce propagation of errors. However, it seems that for a fixed given quantization parameter, if the threshold value is low, the random noise is widely spread. This is due to the fact that the used quantization table in LAR method, that follows the principles described in Eq. 7.4: the higher is QP, the coarser is the quantization on little blocks. As low threshold value leads to large amount of little blocks, the decoded images in those cases seem more deteriorated and lead to more distortions in the synthesized views. Indeed, when $Y = 60$ and $QP = 25$ in both Fig. 7.21 and Fig. 7.22, in the synthesized views, the objects' contours are better rendered than the case of $Y = 1$ and $QP = 25$ because depth values around the edges are better preserved by the quad-tree decomposition. This explains also the density of the errors and its localization on little blocks, as can be seen on the images of the figures 7.21 and 7.22.

To conclude, this study on the influence of the quantization step on synthesized views quality revealed that this bit rate control method can induce noticeable and annoying artifacts on the synthesized views. This effect is enhanced when threshold value Y is low, i. e. when the quad-tree representation contains numerous small blocks.

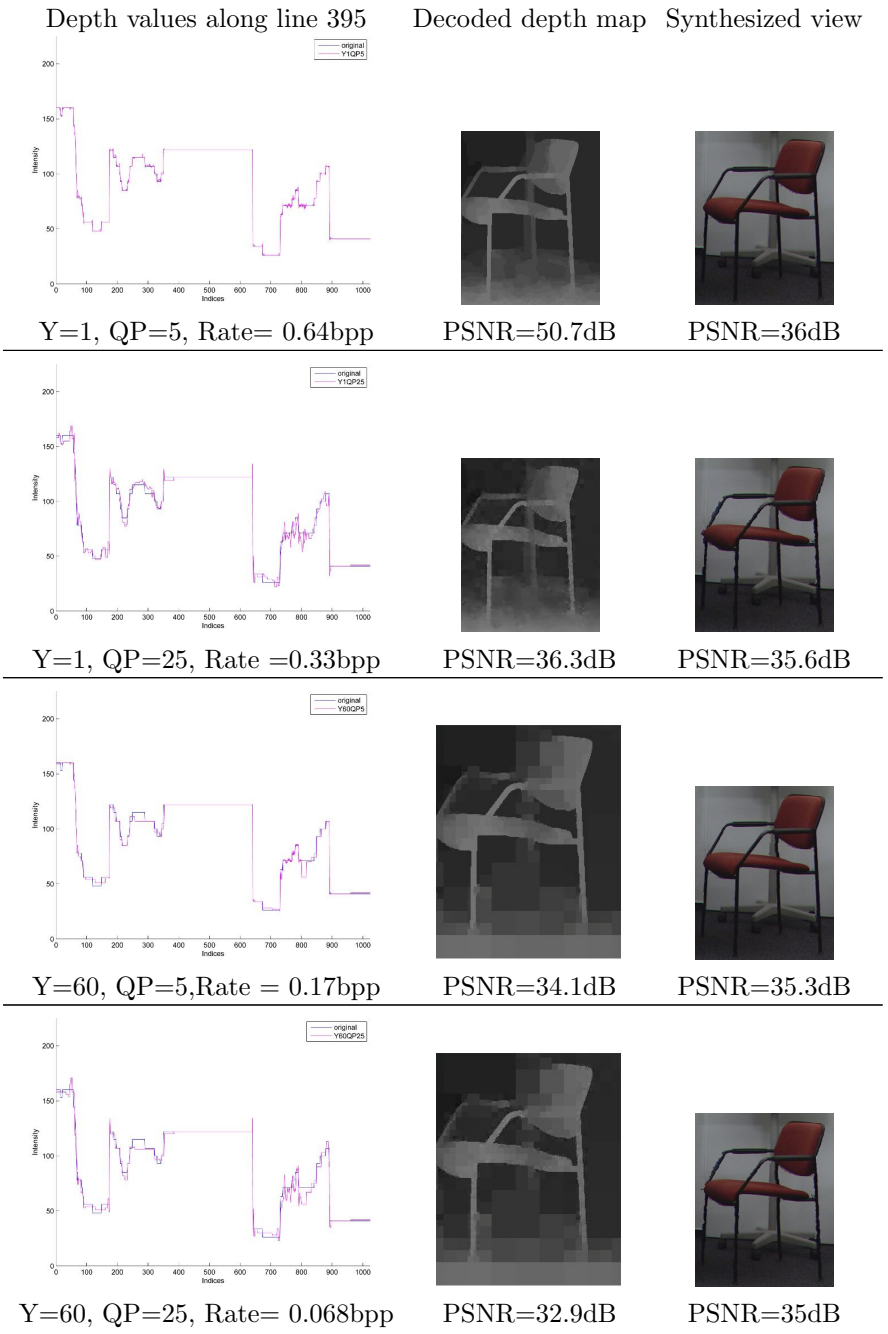


Figure 7.21: *Effect of compression on depth maps and on synthesized views. The first column gives a comparison of values of original and decoded depth maps Book Arrival. The graphs show the depth values along the line number 395 of the image (arbitrarily chosen). The values of the original depth map along the line 395 is depicted in blue and the values of the decompressed depth map along the same line is depicted in pink. The second column gives a snapshot of the decompressed depth map. The third column gives a snapshot of the resulting synthesized frame.*

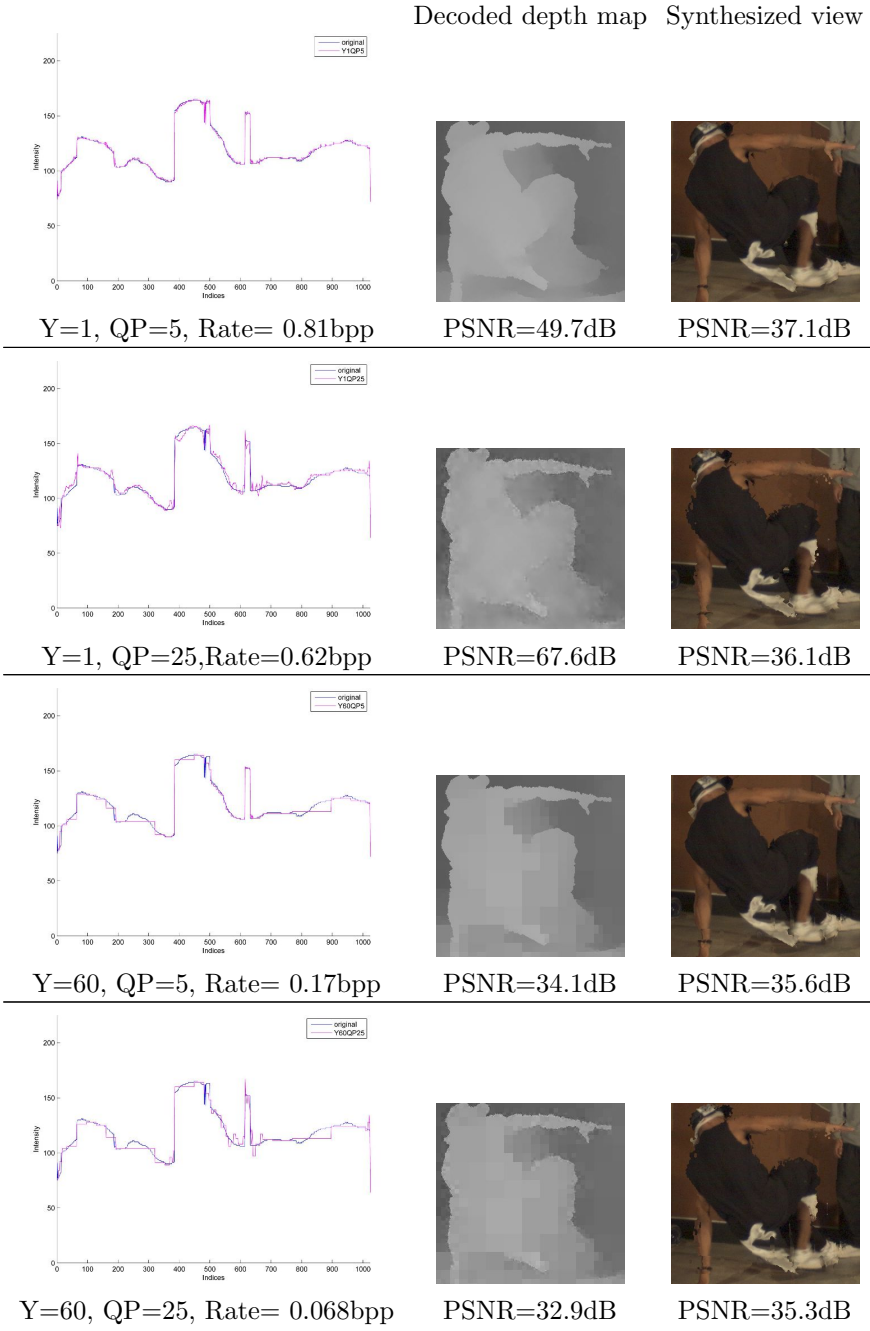


Figure 7.22: *Effect of compression on depth maps and on synthesized views. The first column gives a comparison of values of original and decoded depth maps Breakdancers. The graphs show the depth values along the line number 395 of the image (arbitrarily chosen). The values of the original depth map along the line 395 is depicted in blue and the values of the decompressed depth map along the same line is depicted in pink. The second column gives a snapshot of the decompressed depth map. The third column gives a snapshot of the resulting synthesized frame.*

7.3 Conclusion

This chapter presented the basics of Locally Adapted Resolution coding method and its pyramidal profile extension. Our assumption was that being a content-based encoding method, LAR codec provides tools enabling the respect of depth specificities that contain smooth areas with sharp edges.
The chapter presented studies that investigate the reliability of the first assumption. First, the results showed that the Flat layer of the coder was sufficient for the encoding of depth maps, since the "*Flat + details*" profile induced an additional coding cost without visual quality gain. Second, it appeared that the threshold value Y influences the depth map representation and thus the quality of the rendered view. Although lowering the threshold value Y allows bit-rate savings, this is offset by the fall of visual quality of the synthesized view. Finally, the quantization strategy used in LAR codec seem to be unadapted to depth maps compression. Although the assumption expressed by Eq. 7.4 regarding the influence of block distortion on visual quality is proved for color images, it is not the case for depth maps compression, since small blocks correspond to edges. Yet, edges have to be preserve as most as possible in priority to enable a correct rendered view visual quality.

LAR codec encoding strategies are optimized for color images compression. The next chapters propose new depth maps coding method based on LAR codec and taking into account the conclusions of this first study.

CHAPTER 8

Z-LAR: a new depth map encoding method

The previous chapter evaluated the performances of a 2D codec, namely LAR, in the case of depth map coding. Advantages and drawbacks were discussed thanks to different studies. In particular, the results confirmed our assumption that being a content-based encoding method, LAR codec provides tools enabling the respect of depth specificities that contain smooth areas with sharp edges. Based on these preliminary studies on LAR codec performances, we propose a novel scheme for depth map compression that we call Z-LAR, in this chapter. This study led to the publication of one international conference paper [BMP12a].
The chapter is divided as follows: Sec. 8.1 reminds our goal in the context of MVD coding; Sec. 8.2 describes our main contributions through the design of this new depth coding tool; Sec. 8.3 presents the performances of the proposed scheme by a validation protocol.

8.1 Motivations

Our goal is to achieve depth maps compression because up to now, there is no standardized compression method for MVD sequences. However, MPEG is currently standardizing a novel MVD encoding framework, namely 3DVC, that was previously discussed in Chapter 3. Most of the proposed compression methods rely on the extension of state-of-the-art 2D codecs. The most popular is H264/AVC [STL04] whose 3D extension (standardized for Multi-View-Video representation, MVV), namely H.264/MVC for Multi-view Video Coding [MMSW06], has been the subject of many adaptations for MVD compression [MSD+09]. Previous studies already pointed out the impact of depth encoding on the synthesized frames. Compression-related artifacts that may be imperceptible in depth maps cause important distortion during the synthesis process [MMS+09]. Many methods have been proposed recently in order to address the aforementioned issues. Various encoding strategies are possible to achieve depth map compression. Several studies have proposed bit-rate control methods [MFdW07, DTP09] relying on the objective quality of the resulting synthesized views, or on a distortion model [LHM+09]. A popular and efficient strategy is the post-processing of depth maps after decoding [DSFK+11]. Depth-adapted encoding methods [MdWF06, GRP+10, SD09] have also been proposed. Our work is in line with the depth-adapted encoding strategy since the method proposed in this chapter relies on the content-based representation of the depth map of LAR codec.

Based on the results presented in Chapter 7, we believe that depth map compression can be achieved with LAR codec tools by means of several changes in order to adapt the strategy according to depth maps specificities. The main purpose of this novel framework is to preserve the consistency between color and depth data. Our strategy is motivated by previous studies[MMS+09] of artifacts occurring in synthesized views: most annoying distortions are located around strong depth discontinuities (Chapter 6) and these distortions can be due to misalignment of depth and color edges in decoded images. Thus the method is meant to preserve edges and to ensure consistent localization of color edges and depth edges. The LAR codec is based on a quad-tree representation of the images. In this quad-tree, the smaller the blocks, the higher the probability of the presence of a depth discontinuity. Analogously, big blocks correspond to smooth areas. The quad-tree representation contributes in the preservation of depth transitions when target bit-rate decreases. Another original contribution of the proposed method relies on the use of the decoded color data as an anchor for the enhancement of the associated decoded depth, together with information provided by the quad-tree structure. This is meant to ensure consistency in both types of data after decoding. We also propose to change the quantization strategy so that the artifacts occurring in the rendered view are less perceptible or less annoying.

8.2 Depth map encoding method

Based on the previous results, since depth maps do not contain high frequency areas, the details are not essential and represent an avoidable additional cost of compression. Thus, only the flat image is considered and encoded in the method we propose, i. e. we use "*LAR Flat pyramid only*" profile.

8.2.1 Quad-tree resolution

The quad-tree decomposition is dependent on the local gradient of the depth image. Given a threshold Y for the local gradient, the image is split into blocks: the higher the local activity, the more splits. This leads to small blocks around object edges and bigger ones in continuous areas. In the original LAR method, the minimal size of the blocks, N_{min} is equal 2×2. In previous experiments, we observed that using $N_{min} = 1$ instead of $N_{min} = 2$ provides better visual results on the synthesized view, but increases the bit rate. Since our priority is to enhance the visual quality, we first opt for $N_{min} = 1$.

Pasteau *et al.* [PBD+10] suggested applying a quantization step depending on the block sizes, in the case of conventional images. Our experiments revealed that in the case of depth map compression, this was not an adequate strategy because the smaller the blocks, the coarser was the quantization (this allowed bit rate savings because small block are costly). Yet, small blocks correspond to strong depth discontinuities and errors occurring in these areas may have disastrous effect at the synthesis step. Figure 8.1a) shows the impact of the quantization as suggested in Pasteau *et al.*[PBD+10] (first column) at 0.06 bpp and using $N_{min} = 1$. Depth transitions are highly degraded and will result in errors in the synthesized frame (third column, crumbling artifacts around the head and around the legs of the chair). The synthesized frames obtained in Figure 8.1c) are generated from original color data and decoded depth maps in order to visually assess only the impact of depth quantization (i.e. not the combined effect of both color and depth compression) using Pasteau *et al.* quantization.

a) Decoded depth when applying quantization suggested in Pasteau *et al.*

b) Decoded depth when applying our proposed strategy.

c) Frame synthesized from decoded depth when applying quantization suggested in Pasteau *et al.*

d) Frame synthesized from decoded depth when applying our proposed strategy.

Figure 8.1: *Comparison of two decoded depth maps at 0.06bpp, using the LAR method or the proposed method of rate control.*

Our priority is to preserve the quality of the objects' contours rather than the actual scene depth values. For this reason, we use $N_{min} = 1$ as the minimal size of block in the quad-tree, in order to fit as most as possible the actual contours of the image. Since smallest block of the quad-tree representation are located around the edges, we opt for a spatial prediction of this specific area: the lowest level of LAR pyramid is not encoded to avoid the error propagation and crumbling artifacts observed in the previous chapter. Instead, values of small blocks are predicted according to their neighbor blocks. We thus propose to partly replace hierarchical prediction by spatial prediction. In the following the subsections, we will explain our strategy for bit rate savings.

8.2.2 Truncated pyramid and spatial prediction

The compression scheme in LAR method is based on a pyramidal decomposition [BDR05], previously referred as "*LAR Flat pyramid only*". The pyramid, built from I, consists of a set of images, noted as $\{L_l\}_{l=0}^{l=l_{max}}$, as a multi-resolution representation of the image, where l_{max} is the top of the pyramid and $l = 0$ is the lowest level, i.e. the full resolution image. At each level, the image is obtained by averaging the principal diagonal of each block at previous level, as expressed by Eq. 7.5.

LAR method uses hierarchical prediction that is the prediction of each level of the pyramid, from top to bottom. For each level, the associated image of errors, also relying on the quad-tree decomposition, can be transmitted to compensate the prediction errors. At the decoder side, from the top to the bottom, the image is reconstructed.

Compression cost is mainly due to the encoding of small blocks. As explained, small blocks are not transmitted (those are blocks whose size is such as $N = N_{min}$). This is achievable thanks to the pyramidal decomposition. The encoding of small blocks is related to the image of errors corresponding to the lowest level, i.e. L_0. The lowest level is not encoded in the method we propose, and the image will be refined at the decoder side thanks to the analysis of the values of the nearest neighbor blocks whose size is such as $N > N_{min}$: they will be predicted, depending on the values of their closest larger blocks. This allows bit-rate savings. The pseudo code of this prediction is given in Algorithm 8.1.

8.2.3 Spatial quantization of depth

As explained in the previous subsections, we opt for a quad-tree decomposition such as the minimal block size is $N = N_{min}$. Although we believe that this choice allows better visual performances on the synthesized views, small blocks are costly and we need a bit rate control strategy. We propose a quantization achieved through the evolution of the quad-tree representation of the image. Small blocks are costly and a way to reduce the bit-rate is to reduce the number of small blocks. This implies that the quad-tree representation can change according to the target bit-rate. The number of small blocks is directly related to the value of the threshold Y. Thus, an increasing threshold Y decreases the bit-rate, so that the representation of the image contains larger blocks. This corresponds to a spatial quantization that concerns depth values. It results in assigning the same depth value to objects that were not formerly in the same depth plane. The dynamic range of depth is thus reduced but the global structure is preserved. Fig. 8.2 gives the quad-tree representations and the resulting depth maps using two different thresholds for the quad-tree decomposition. It shows that the semantic information of the image is preserved. Fig. 8.1b) shows that the proposed method (second column) renders sharp depth transitions.

Algorithm 8.1: Prediction of full resolution image from truncated pyramid

Require: $\tilde{L}_l$ is the estimated representation of the image at the decoder side, for level l,
Quad-tree$^{[N_{max}...N_{min}]}$ is the quad-tree partition.
Decoding of truncated pyramid:
for $l = l_{max} \dots l_1$ **do**
  Estimate $\tilde{L}_l$ as in the LAR method
end for
for each block of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N = N_{max} \dots N_{min}$ **do**
  Given **Quad-tree**$^{[N_{max}...N_{min}]}$, then $\tilde{L}_0(b^N(i,j)) = \tilde{L}_1$
end for
Decoding of full resolution image:
for each block of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N = N_{min}$ **do**
  $\tilde{L}_0(b^(N_{min})(i,j))$ = Mean value of the closest block b^N of **Quad-tree**$^{[N_{max}...N_{min}]}$ such as $N > N_{min}$
end for
return $\tilde{L}_0$

The synthesized frame in Fig. 8.1d), fourth column, shows improvements compared to the previous strategy, third column (Fig. 8.1c)).

8.2.4 Smooth depth reduction with rate

The threshold value Y can vary from 1 to 254. This leads to a specificity of this depth coding strategy. The higher the threshold value Y, the more the depth structure is modified and diminished. Indeed, the higher the threshold value Y, the more large blocks in the quad-tree partition. The final reconstructed depth map corresponds to the average values of each block of the partition. Thus this coding strategy results in a uniform depth map for $Y \simeq 254$ (very low bit-rate). Then, at the synthesis step, such a uniform depth map leads to the projection of every color pixel to the same depth. In other words, the resulting synthesized view, observed in monoscopic conditions, is a 2D image. So the lower the bit-rate, the less depth in the scene. So a new type of artifact is introduced by the proposed method: reduction of depth. This is illustrated by Fig. 8.3. This figure shows the depth values of pixels belonging to a considered line of the depth map (line 250, in red on Fig. 8.3(a)). The depth structure of the scene decreases while Y increases, until it is completely flat. So for $Y = 201$ in this example, all color pixels are assigned to the same depth.

8.2.5 Depth reconstruction at decoder side

On main requirement for visual quality is the consistency between color and depth edges. A reconstruction step is included at the depth map decoder side, right after the first estimation of the smallest blocks. To enforce the color/depth edges consistency, a second pass of depth reconstruction was introduced. The proposed feature consists of a multi-lateral filtering aided by the quad-tree representation whose principles are partially based on the literature. Numerous studies already addressed the issue of depth map accuracy by depth filtering processes [MLD12, DSFK+11, LTL10]. Min *et al.* [MLD12] proposed a weighted filtering based on the analysis of a color-aided joint histogram. De Silva *et al.* [DSFK+11] also proposed an analysis of the decompressed depth maps' histograms for the filtering stage. Lai *et al.* [LTL10] apply a multi-lateral filter on the whole depth

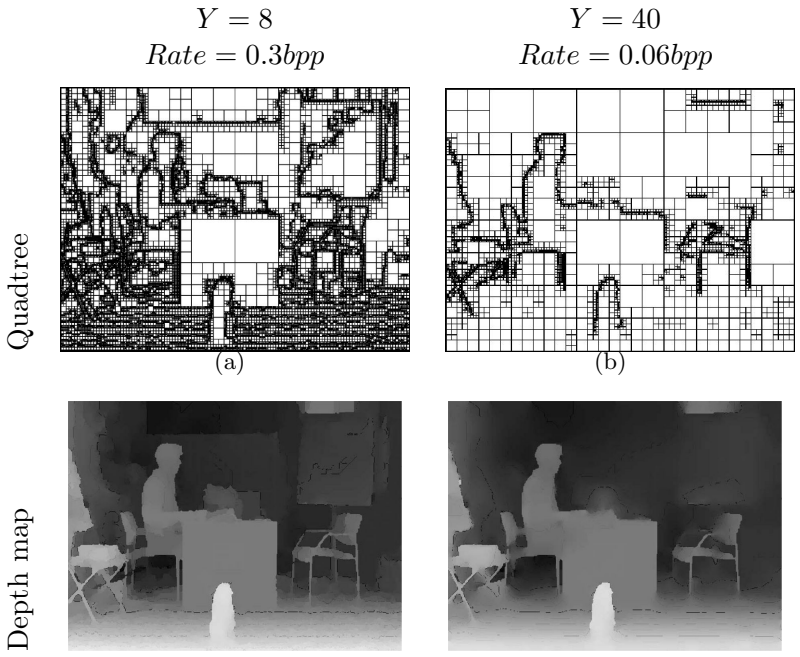


Figure 8.2: *Quantization of the depth map.*

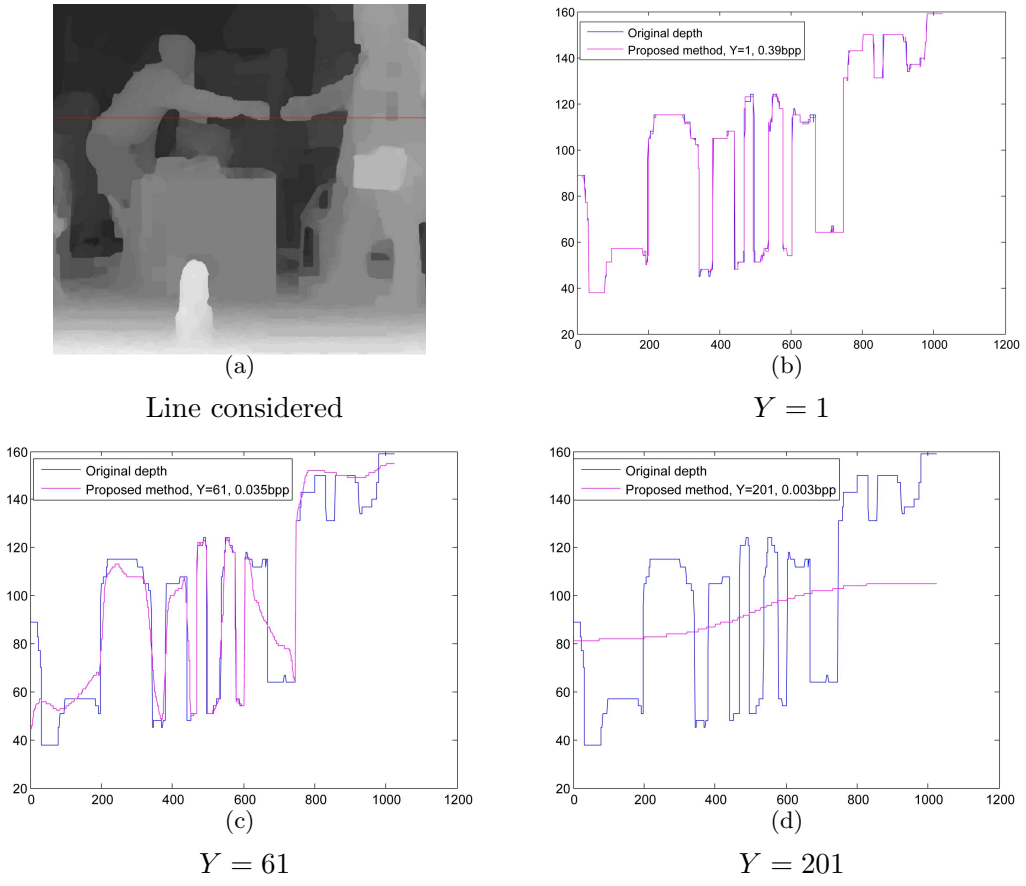


Figure 8.3: *Depth values along line 250 of frame 33, Book Arrival sequence, view 6.*

map, relying on the associated original color image. In our proposed method, the decoded associated color image is used to enhance only the blocks smaller or equal to 4×4 in the depth map. This color image can be encoded by any state-of-the-art color codec method since our contribution targets depth coding only. Thanks to the decoded quad-tree decomposition from the decoded LAR stream, the location of small blocks is already known. Small blocks are likely to be located around depth discontinuities. Thus, it is believed that improving the accuracy in these regions, according to the decoded associated color, will ensure consistency between color and depth edges. Our filter is thus only applied on small blocks. Let $\tilde{C}$ be the decoded associated color image, and $\tilde{L}_0$ the lowest level image of the depth pyramidal decomposition. Let Ω be the set of pixels of the quad-tree partition **Quad-tree**$^{[N_{max}...N_{min}]}$ whose size N lies in $N \in [N_{min} \dots 4]$, such as:

$$\Omega = \left. \tilde{L}_0(x,y) \right| \tilde{L}_0\left(x,y\right) \in \tilde{L}_0\left(b^N\left(i,j\right)\right), \quad N \in [N_{min} \dots 4] \tag{8.1}$$

Ω thus represent the set of so-called small blocks.
The reconstruction, noted $\tilde{L}_{0r}\left(x,y\right)$, of any pixel belonging to Ω is expressed as:

$\forall\ \tilde{L}_0\left(x,y\right) \in \Omega,$

$$\tilde{L}_{0r}\left(x,y\right) = \tilde{L}_{0r}\left(p\right) = \frac{1}{K} \sum_{q \in \Gamma} \tilde{L}_0\left(p\right) e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|\tilde{L}_0(p)-\tilde{L}_0(q)\|}{2\sigma_s}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}} \tag{8.2}$$

$$K = \sum_{q \in \Gamma} e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|\tilde{L}_0(p)-\tilde{L}_0(q)\|}{2\sigma_s}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}} \tag{8.3}$$

Γ is the pixel window used for the calculation; $Luma$ is the luminance component of the decoded color image; $Luma(p)$ and $Luma(q)$ are pixels of the luminance component of the decoded color image; σ_d, σ_s, σ_c are standard deviations related to the spatial domain, the depth range domain (similarity of depth values), and the color range domain, respectively.

Figure 8.4 gives the overview of the proposed method. In this figure, at the encoding step, black blocks correspond to non transmitted blocks.

8.3 Experiments

8.3.1 Protocol

The proposed method is compared to state-of-the-art codec H.264 in intra mode. The choice for this method in this experiment is motivated by the fact that it is a reference method which is usually used as anchor in standardization process. As preliminary studies, experiments concern only still images. First frames of views 6 and 10 from *Book Arrival* were encoded through both encoding methods. Afterwards, decoded color and depth maps were used to compute the intermediate view 8, through the reference software, VSRS 3.5 [TFS+08]. Since view 8 is among the originally acquired views, it is considered as a ground truth for quality assessment. In this paper, the quad-tree decomposition parameters are $N_{min} = 1$ and $N_{max} = 12$. In Equation 8.2, $\sigma_d = 4$, $\sigma_s = 10$, $\sigma_c = 3$. The color images are encoded with H.264, QP varying from 0 to 50. Figure 8.5 gives an overview of the experimental protocol.

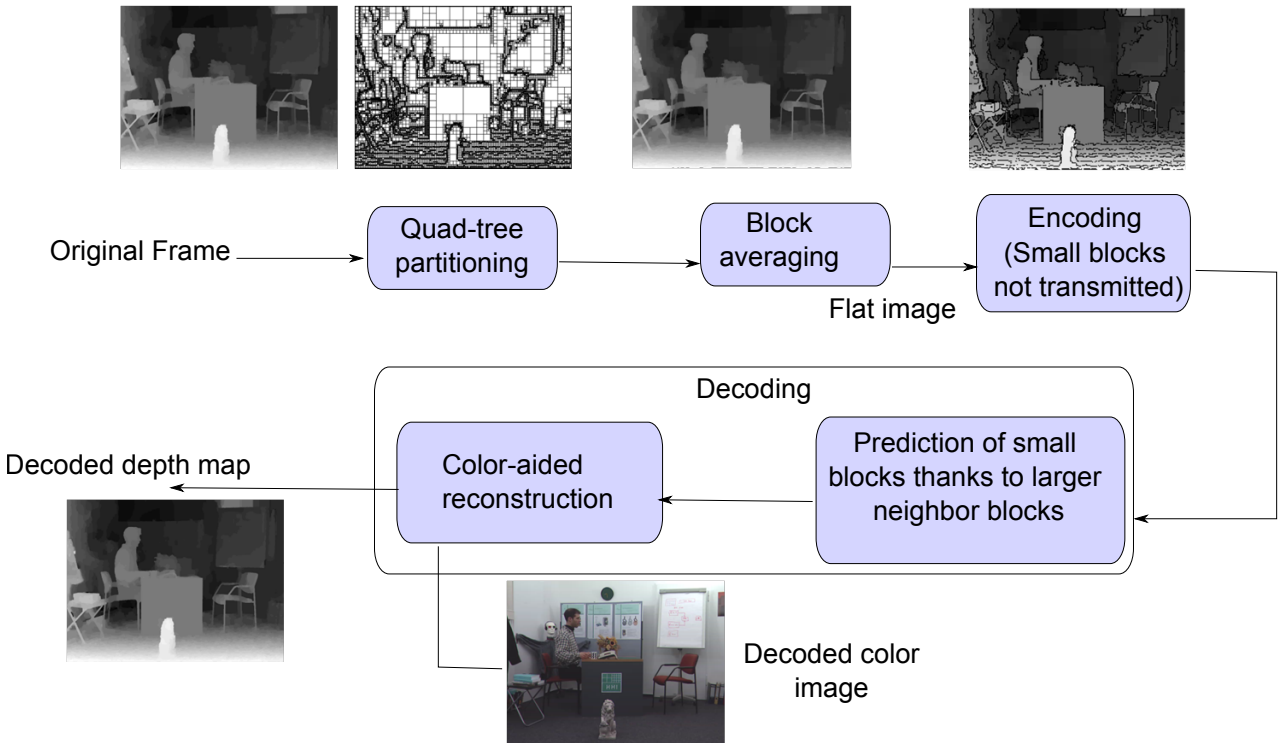


Figure 8.4: *Overview of Z-LAR method.*

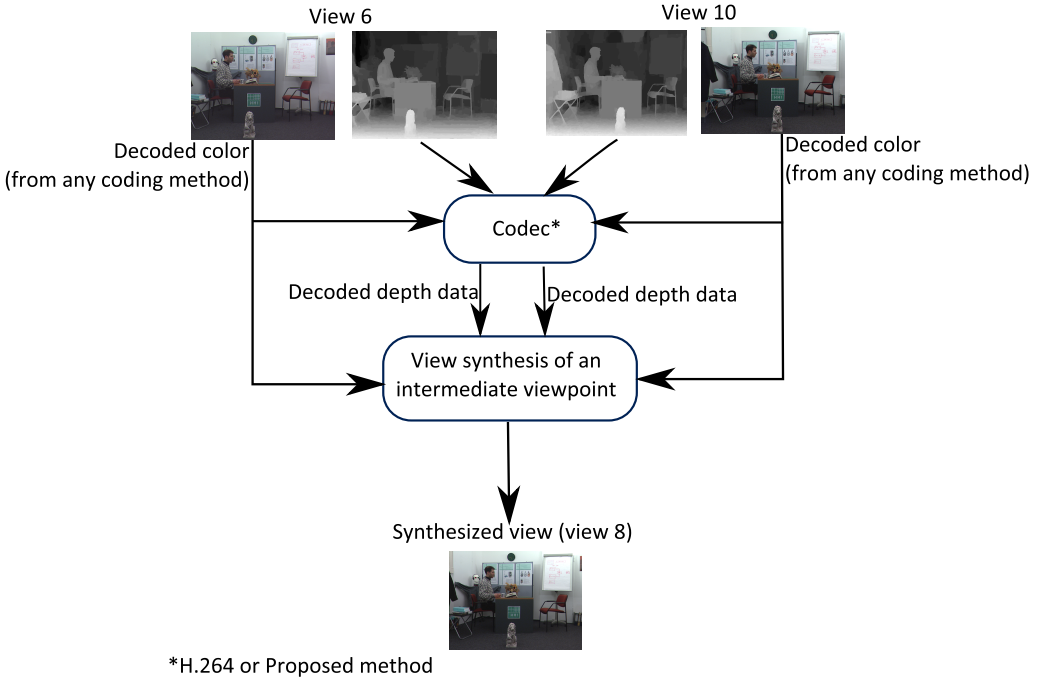


Figure 8.5: *Overview of the experimental protocol*

8.3.2 Results

For the performance comparisons, we ran the objective evaluations with the same set of metrics as those tested in Chapter 6. Since the results output by the objective metrics were higly correlated, we only display here a pixel-based metric (PSNR) and a more perception-oriented metric VIF (Visual Information Fidelity [SB06]) are considered (note that PSNR HVS, VIFP and SNR metrics gave results similar to that of VIF in this experiment). Figure 8.6 depicts the rate-distortion curve obtained by computing PSNR scores and VIF scores of the synthesized views, with respect to the original acquired view. At high bit-rates (higher than 2bpp), the proposed method obtains better PSNR scores (the maximum difference in PSNR is 0.043dB at high bit-rate). However, under 2bpp, H.264 obtains better PSNR scores (the maximum difference in PSNR is 0.37dB at high bit-rate).

The curve based on VIF scores shows that H.264 and the proposed method give similar results at high bit-rates (higher than 2bpp). However, contrary to the curve based on PSNR, the curve based on the perception-oriented VIF shows that the proposed method performs better at low bit-rates (the maximum improvement reaches 5.68%).

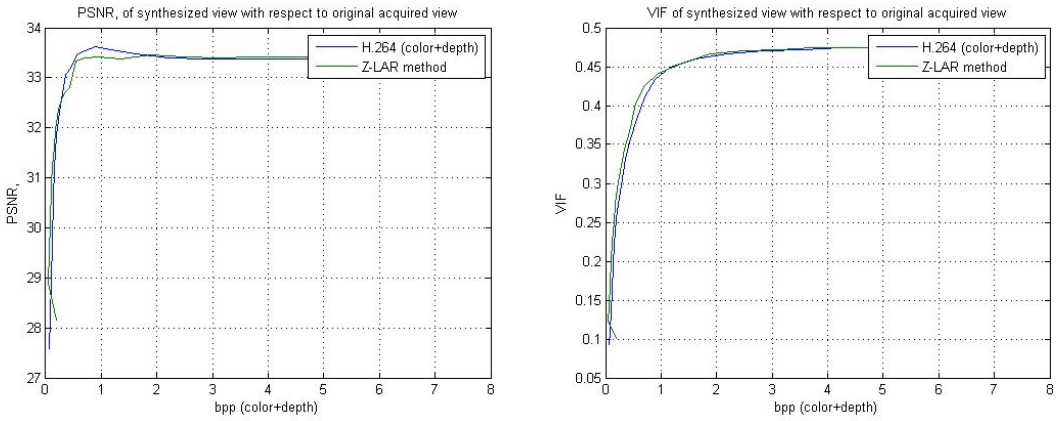


Figure 8.6: *Performance comparisons, in terms of PSNR and VIF, between the original view and the synthesized view.*

A visual appreciation is thus useful to evaluate the methods. Figure 8.7 gives snapshots of the obtained synthesized views for 0.1bpp and 0.9bpp. Ghosting effect is perceptible with both methods behind the head of the man for 0.1bpp. Z-LAR method preserves better the vertical edges: the vertical dark lines of the posters are better rendered with the data encoded with Z-LAR method. At low bit-rate (0.1bpp), Figure 8.7 gives snapshots of the synthesized views. Although, PSNR score shows lower performances for Z-LAR at low bit-rate, the observation of Figure 8.7 shows improvements around the edges of the synthesized objects. The ghosting effect around the head of the man is less perceptible with Z-LAR method. The crumbling artifacts occurring around the leg of the chair at 0.1bpp with H.264 are no longer perceptible with the proposed method.

8.4 Conclusion

We proposed a novel depth map coding framework called Z-LAR whose main purpose is to preserve consistency between color and depth edges to improve the synthesized views

Figure 8.7: *Snapshots of synthesized views from data encoded with H.264 and from data encoded with the proposed method.*

quality. Depth encoding is based on LAR codec. It consists in a quad-tree representation of the images. The quad-tree representation contributes in the preservation of edges in depth data. Our contributions can be listed as follows:

- we opted for a higher quad-tree resolution (smallest block size is $N_{min} = 1$ instead of $N_{min} = 2$), in order to preserve as most as possible the contours and to improve the rendering of the virtual views,
- we proposed a depth spatial quantization: the actual depth structure of the scene is modified but the objects' contours are better rendered than state-of-the-art methods such as H.264. This depth quantization strategy gets rid of "crumbling" artifacts because it does not target the quantization of critical areas corresponding to depth discontinuities.
- we included a second pass for depth reconstruction that makes use of decoded color data as an anchor for the associated depth enhancement at the decoder side.

The originality of our contribution lies on the possibility to obtain a 2D image for very low bit rates, thanks to the spatial quantization we proposed. Our choices were also motivated by the fact that recent studies showed that observers prefer not to have depth rather than quantization artifacts [BCLC09]. The proposed method showed visual performances similar to H.264 at high bit-rates and some improvements at lower bit-rates because it

better preserves object edges. The next chapter will present a second extension of this depth map coding method.

CHAPTER 9

Z-LAR-RP: hierarchical region-based prediction in Z-LAR

In this chapter, we propose a second LAR-based approach for depth maps compression called Z-LAR-RP. This method is meant to be more reliable and scalable because it exploits the pyramidal images to allow multiple depth maps resolution. The prediction technique is based on the region segmentation relying on the decoded quad-tree only, as extracted from the LAR stream. Sec. 9.1 gives an overview of this second contribution to depth map coding. Sec. 9.2 details the tools included in Z-LAR-RP. Sec. 9.3 discusses the performances through quality assessment measurements of synthesized views.

9.1 Overview

Compared to Z-LAR method, the method that will be presented in the following, namely Z-LAR-RP, differs on the prediction step. The minimal quad-tree block size is kept as 1×1.
The associated texture view can be encoded by any state-of-the-art color codec. The Z-LAR-RP uses the decompressed texture information to improve the prediction step involved in depth maps decoding. The compression scheme still relies on the pyramidal profile of LAR, previously referred as "*LAR Flat pyramid only*". In the previous proposed approach, the selection of the lowest level to be transmitted and decoded from the pyramid construction was not allowed. Yet, this option is available in the pyramidal profile of LAR codec for 2D color images. The method that will be presented is meant to overcome this limitation and to allow the selection of depth resolution and mainly to increase the performances of the depth map coding framework (in terms of visual quality of the synthesized views and in terms of complexity).
Any level of the pyramid can be chosen as the lowest to be transmitted and the actual depth map size is reached thanks to a region-based prediction method, that will be presented in the following. Regarding rate control strategy and quantization, this method follows the same principle as Z-LAR, in Sec. 8.2.3: the actual depth structure of the scene is modified when the bit-rate decreases. Fig. 9.1 gives an overview of the method.
In the following section, we show that the basic region-based segmentation method can be jointly used with decoded color data in order to improve the prediction step by propagating the decoded depth values in the smallest blocks of the quad-tree. Then, validation experiments show the performances of the proposed Z-LAR-RP method.

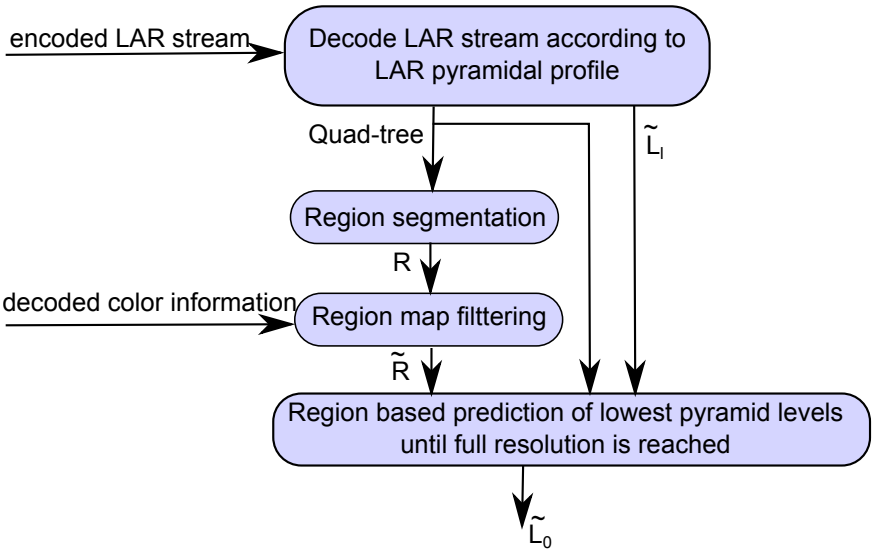


Figure 9.1: *Overview of the Z-LAR-RP*

9.2 Depth map encoding method

9.2.1 Region segmentation from decoded quad-tree

The region segmentation algorithm employed in this method, relies on previous work by C. Strauss presented in [Str11]. The specificity of this segmentation algorithm is that it only requires the knowledge of the image structure, that is contained in the quad-tree partioning of the image as input data. This quad-tree partioning is embedded in the very beginning of the LAR codec bitstream and can be extracted at the decoder side. Algorithm 9.1 gives the details of the segmentation algorithm as described in [Str11].

After creating the seeds from the larger blocks, adjacent regions are agglomerated by region growing. The process is reiterated $iter(CurrentSurf$ times. The number of growing iterations increases with the decrease of the size threshold $CurrentSurf$. Along the merging process, the number of regions should thus decrease. Fig. 9.2 depicts an example of the segmentation result. Fig. 9.2(a) is the quad-tree partition obtained from the first frame of *Breakdancers* depth map of camera 0, with $Y = 5$. Fig. 9.2(b) gives the first seeds, from the larger blocks. Fig. 9.2(c) gives the final region segmentation, with 370 regions.

9.2.2 Color-consistent region edge refinement

In order to enhance segmentation results, we introduce the color information of the corresponding decoded color view. A discrete bilateral filter is applied on the region map obtained from the region segmentation process in order to refine the location of decoded depth map edges to be consistent with color map edges. Algorithm 9.2 gives the details of the method. The region map is denoted R. Any pixel p at location (i, j) belongs to labeled region $R(p) = R(i, j)$ in region map. The filtered region map is noted as $\tilde{R}$.
For each pixel p, a support Γ_p is considered, that is the neighborhood of$p$, centered on p. The filter proceeds in way that Pixel p will be given the most likely region label according to the importance of its neighbors. This importance (or weight) of each neighbor is evaluated regarding its color similarity with p in the corresponding location in the decoded

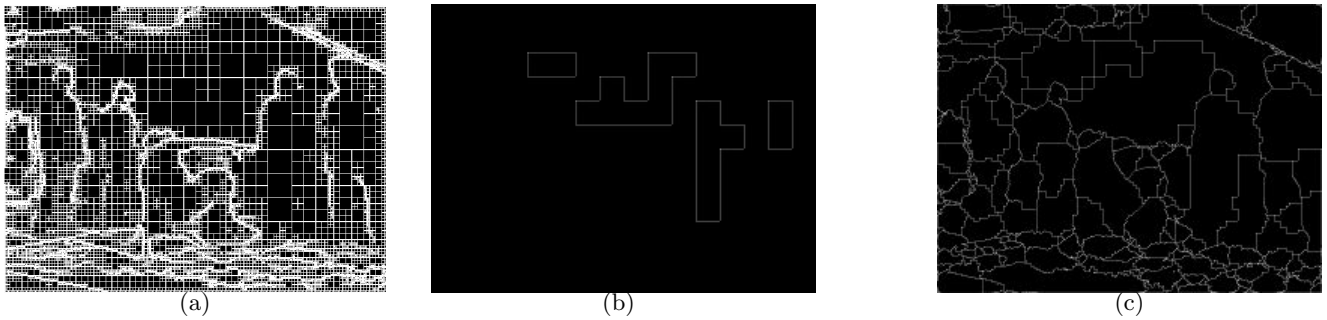
Algorithm 9.1: Region segmentation algorithm from [Str11]

Require: **Quad-tree**$^{[N_{max}...N_{min}]}$ the dyadic quad-tree partition containing P square blocks b_i, $i \in \{1 \dots P\}$ where each block b_i has a surface of $2^S \times 2^S$ pixels,
$S \in \{1 \dots N_{max}$;
Δ^k is the region map after k merging steps;
R_i^k in Δ^k is the k non overlapping region label;
$surf(R_i^k)$ is the surface in pixels of R_i^k;
A_i^k is the set of adjacent regions of R_i^k in Δ^k.
Initializations
$k = 0$
$\Delta^k =$ **Quad-tree**$^{[N_{max}...N_{min}]}$
$CurrentSurf = 2^{N_{max}} \times 2^{N_{max}}$
repeat
  Seeds creation
  while $\exists R_i^k | surf(R_i^k) = CurrentSurf$ **do**
    while $\exists R_j^k \in A_i^k$ and $surf(R_j^k) = CurrentSurf$ **do**
      Merge R_j^k and R_i^k into Δ^{k+1}
      $k = k + 1$
      Update A_i^k
    end while
  end while
  $CurrentSurf = \lfloor CurrentSurf/4 \rfloor$ {Region growing}
  while $\exists R_i^k | surf(R_i^k) = CurrentSurf$ **do**
    for iter=1 to $iter(CurrentSurf)$ **do**
      Let A' = $\{R_j^0 | R_j^0 \in A_i^k$ and $surf(R_j^0) = CurrentSurf$
      Let Z = card(A')
      Merge R_i^k and A' into Δ^{k+Z}
      $k = k + Z$
      Update A_i^k
    end for
  end while
until $CurrentSurf = 0$

color image, and regarding its distance to p. Finally p is allocated the same region label as the neighbor having the highest importance (or weight)
The luminance component *Luma* of the decoded texture view is used to estimate the color similarity of the considered neighborhood. The algorithm 9.2 aims at assigning each pixel of the region map the more likely region label according to the criterion described earlier. These constraints are are expressed by the factors σ_c and σ_d respectively.
Fig. 9.3 depicts a snapshot of the result of this process over the region map, obtained with a neighborhood of 7×7 pixels, centered on the processed pixel, $\sigma_c = 30$ and $\sigma_d = 3$. The region frontiers in white are superimposed on the original corresponding color view. It can be observed that the segmentation is more consistent to color data.

9.2.3 Pyramid truncation

Any level l of the pyramid can be chosen as the lowest to be transmitted and the actual depth map size is reached thanks to the region-based prediction method described by



(a) Quad-tree partition, $Y = 5$ (b) First seeds (c) Region segmentation result with 370 regio

Figure 9.2: *Region segmentation using [Str11]*

Algorithm 9.2: Region segmentation enhancement of the depth map based on the decoded color information

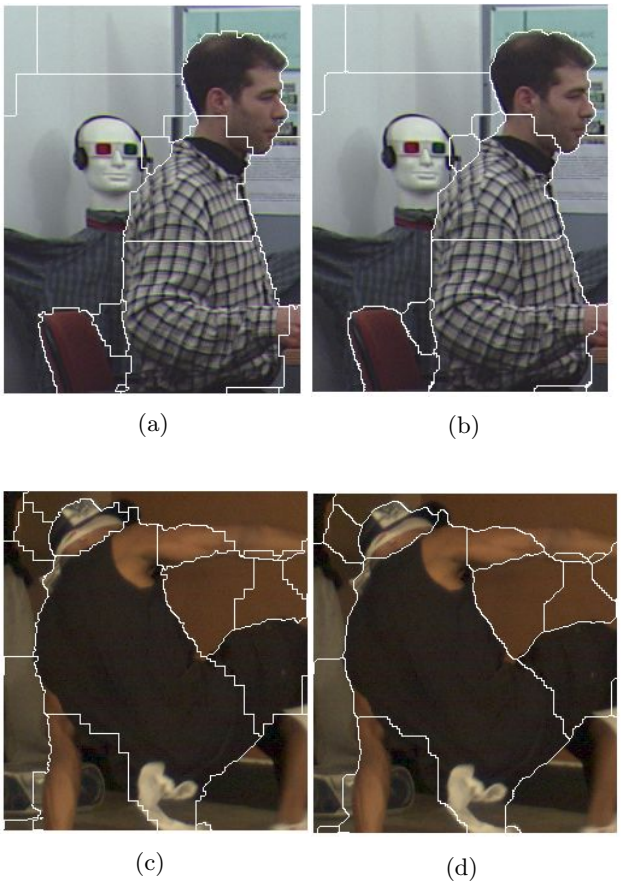
Require: R the region map of the depth image with $N_{regions}$ labels;
$W[N_{regions}]$ the array of region weights;
$Luma$ the associated decoded texture image
Initializations
$Temp(p) = Temp(i,j) = R(p) = R(i,j) \mid \{p = (i,j) \in N_x \times N_y\}$
$W[k] = 0 | \{k \in [1 \dots N_{regions}]\}$
for all $p \in R$ **do**
  for all $q \in \Gamma_p$ **do**
    $r = R(q)$
    $W[r] = W[r] + e^{-\frac{\|p-q\|}{2\sigma_d}} e^{-\frac{\|Luma(p)-Luma(q)\|}{2\sigma_c}}$
  end for
  Find $\tilde{r} \mid \tilde{r} = \underset{k \in [1 \dots N_{regions}]}{\text{argmax}} W[k]$
  $Temp(p) = \tilde{r}$
  Reset all elements of W to 0
end for
$\tilde{R}(i,j) = Temp(i,j) \mid \{(i,j) \in N_x \times N_y\}$
return $\tilde{R}$

Algorithm 9.3. Any pixel of coordinates (i,j) is denoted as p. $\tilde{L}_{l_{min}}$ is the lowest encoded level image of the pyramid, with $l_{min} \geq 1$. The block $b^N(i,j)$ is as described in Eq. 7.1: $b^N(i,j)$ is a block of size N, located at (i,j) in the quad-tree partition.N is the block size as described in Eq. 7.2.For each predicted pixel p in the magnified level, a support $\Gamma_{\lfloor \frac{p}{2} \rfloor}$ is considered, that is the pixel neighborhood in $\tilde{L}_{l_{min}}$, the lowest decoded level image, centered on the corresponding processed pixel $\lfloor \frac{p}{2} \rfloor$. K is a normalizing factor defined as:

$$K = \sum_{q \in \Gamma_{\lfloor \frac{p}{2} \rfloor}} \delta_p(q) \cdot e^{-\frac{\|\lfloor \frac{p}{2} \rfloor - q\|}{2\sigma_1}} \cdot e^{-\frac{\|\tilde{L}_l(\lfloor \frac{p}{2} \rfloor) - \tilde{L}_l(q)\|}{2\sigma_2}}, \tag{9.1}$$

where $\delta_p(q)$ is the existence function defined as:

$$\delta_p(q) = \begin{cases} 1 & \text{if } \tilde{R}(p) = \tilde{R}(q) \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$



(a) (b)

(c) (d)

Result of region segmentation

Region segmentation after enhancement process

Figure 9.3: *Region segmentation after applying enhancement process*

The reconstruction of depth lowest level images is based on a weighting sum of the corresponding neighbors in the direct upper level of the pyramid. The neighbors contribute into this weighting sum only if the belong to the same region in the full image resolution.

Algorithm 9.3: Region-based depth map prediction

Require: $\tilde{L_{l_{min}}}$ the depth lowest decoded level image of the depth map LAR pyramid with $l_{min} \geq 1$;
Quad-tree$^{[N_{max}...N_{min}]}$ the quad-tree partition;
$\tilde{R}$ the filtered region map.

repeat
  for all $p \in \tilde{L}_{l-1}$ **do**
    if $\tilde{L}_{l-1} \in b^N \mid N < 2^l$ **then**
      $\tilde{L}_{l-1}(p) = \frac{1}{K} \sum_{q \in \Gamma_{\lfloor \frac{p}{2} \rfloor}} \tilde{L}_l(q) \cdot \delta_p(q) \cdot e^{-\frac{\| \lfloor \frac{p}{2} \rfloor - q \|}{2\sigma_1}} \cdot e^{-\frac{\| \tilde{L}_l(\lfloor \frac{p}{2} \rfloor) - \tilde{L}_l(q) \|}{2\sigma_2}}$
    else
      $\tilde{L}_{l-1}(p) = \tilde{L}_l(\lfloor \frac{p}{2} \rfloor)$
    end if
  end for
until $l = 0$

9.3 Experiment 1: objective quality assessment

9.3.1 Experimental protocol

The goal of these experiments is the validation of the Z-LAR-RP as an alternative to depth map coding. So only depth maps are encoded in order to highlight the impact of depth quantization strategies. Fig. 9.13 depicts the general scheme followed in these experiments. Depth coder under tests include the Z-LAR-RP, HEVC 6.1 and H.264 (JM 18) both in intra coding mode. The choice for these methods in this experiment is motivated by the fact that they are reference methods which are usually used as anchors in standardization process. Table 9.4 gives the details of the quantization parameters used in these experiments. Six MVD sequences are used in these experiments: *Book Arrival*, *Newspaper*, *Kendo*, *Balloons* are real scenes; and *GT_Fly* and *Undo_Dancer* are synthetic scenes. Table 9.1 summarizes the sequences' features. The sequences were selected for their availability and amount of depth. The key frames were selected for their amount of depth. Table 9.5 gives the details of the encoded viewpoints and the target viewpoint for the synthesis. The synthesis process is performed throught the very last release of VSRS, that is the version used in MPEG 3DV group of standardization at the time of writing this thesis.

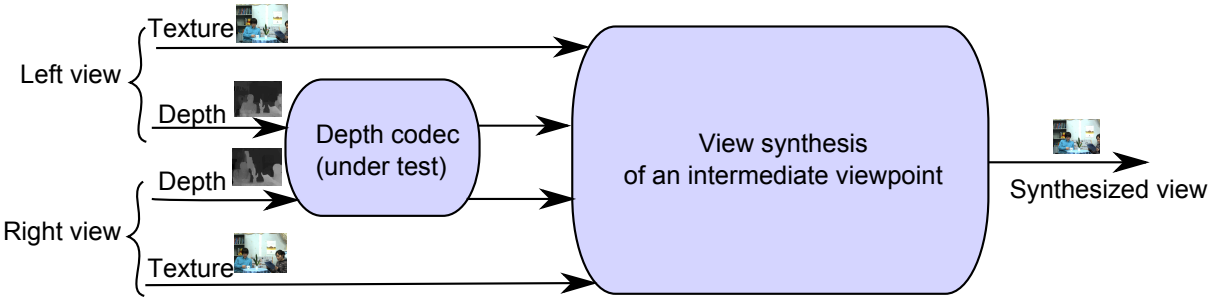


Figure 9.4: *Overview of the experimental protocol.*

Sequence Name	Resolution	No. of frames	Camera Arrangement
Book Arrival	1024×768	100	16 cameras with 6.5cm spacing
Newspaper	1024×768	300	9 cameras with 5 cm spacing
Balloons	1024×768	300	7 cameras with 5 cm spacing, moving camera array
Kendo	1024×768	300	7 cameras with 5 cm spacing, moving camera array
GT_Fly	1920×1080	250	Computer generated imagery with ground truth depth data
Undo_Dancer	1920×1080	250	Computer generated imagery with ground truth depth data

Table 9.1: *Six MVD sequences used in the experiments.*

Sequence Name	Encoded view points	View to synthesize	Frame no.
Book Arrival	$10 - 6$	8	33
Newspaper	$2 - 6$	4	1
Balloons	$1 - 5$	3	1
Kendo	$1 - 5$	3	1
GT_Fly	$1 - 9$	5	157
Undo_Dancer	$1 - 9$	5	250

Table 9.2: *Input and output views of the experiment.*

Depth codec	Quantization parameter
H.264 (JM18)	Qp = [25, 27, 30, 33, 35, 37, 40, 42, 45, 47]
HEVC 6.1	Qp = [34, 36, 39, 41, 42, 43, 45, 46, 48, 50]
Z-LAR-RP	$Y = \{1$ to 241 $\}$, step by 10

Table 9.3: *Input and output views of the experiment.*

9.3.2 Results

Fig. 9.5 and Fig. 9.6 depict the results of objective assessments through the widely used PSNR and MSSIM. However, since it has be shown in Chapter 6, the objective metrics are not sufficient to predict human perception of synthesized views quality, though MSSIM was one of the objective metrics giving the best results out of the tested set of metrics. Moreover, in the case of our proposed coding scheme, objective measurements based on the fidelity such as PSNR and MSSIM are inappropriate. Indeed, our coding method modifies the depth structure of the scene. Thus objects may be shifted. Since objective metrics are mostly FR, they measure the fidelity between two images and it is expected that our method obtain bad scores while having good visual quality performances. So we provide the PSNR of depth maps (average between the two views), the PSNR of the synthesized view, with the original acquired view as the reference and the MSSIM of the synthesized view, with the original acquired view as the reference, in Fig. 9.5 (for *Balloons*, *Kendo* and *Book Arrival*) and Fig. 9.6 (for *Newspaper*, *GT_Fly* and *Undo_Dancer*), both as a rough guide. Snapshots of the corresponding views are provided in Fig. 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12. Note that it can be observed a slight shift for Z-LAR-RP snapshots. The same viewpoint is always generated but at very low bit-rates, Z-LAR-RP tends to deliver a uniform depth map which results in a slight shift of the scene in the synthesized view. As expected the objective measures rate the Z-LAR-RP as worst than the two state-of-the-art codecs. This was expected because of the reasons mentioned above. However, visual analysis of all the synthesized views proves that the quality is often similar (Fig. 9.11) or even superior than that of state-of-the-art methods (Fig. 9.7, 9.8, 9.9, 9.10, 9.12). Moreover the proposed scheme allows very low bit-rates (around 0.003bpp). In these cases, the proposed scheme automatically transmit a flat depth map, which results in a good visual rendered view quality.

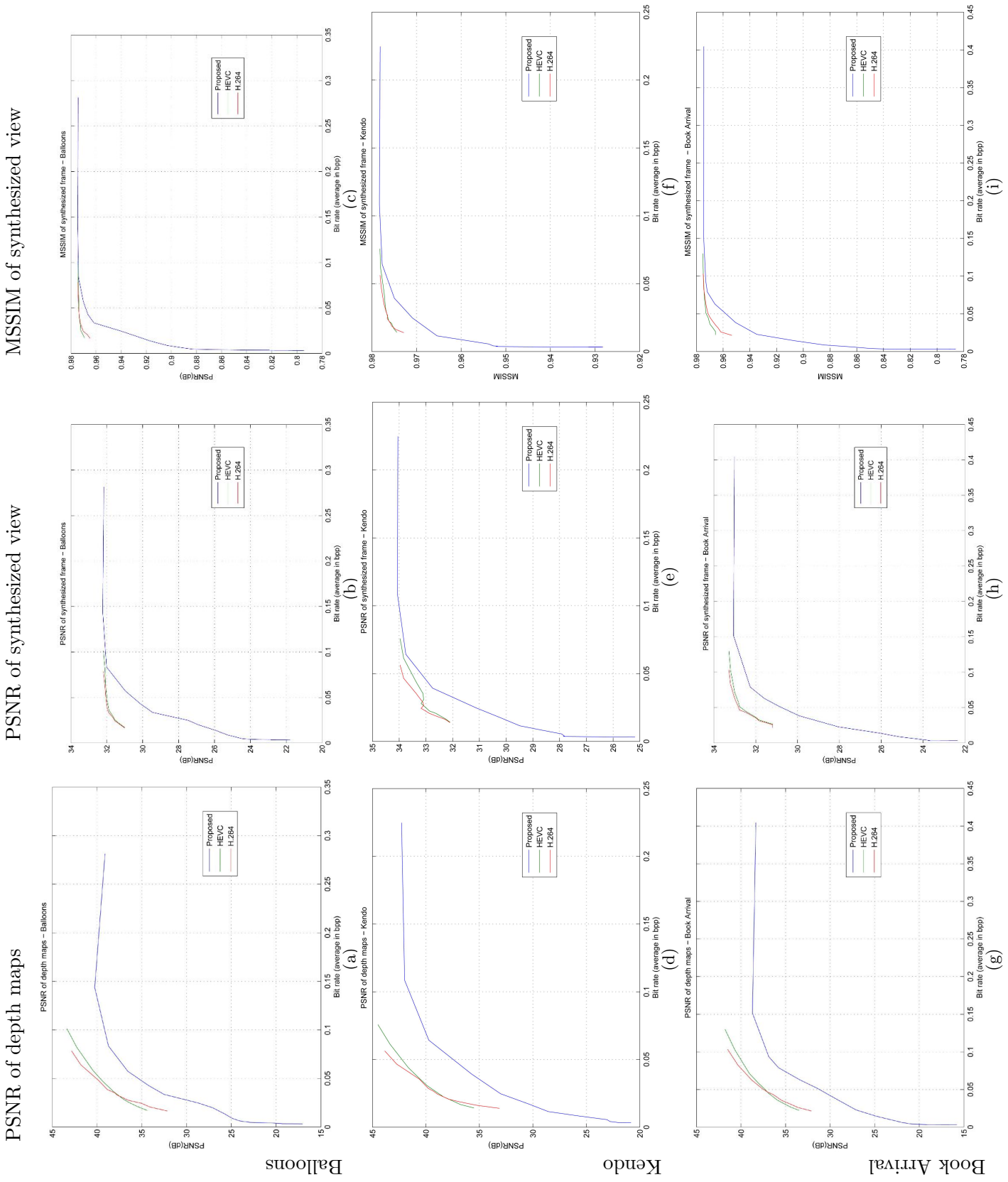


Figure 9.5: *Rate/distortion curves of depth maps and synthesized views.*

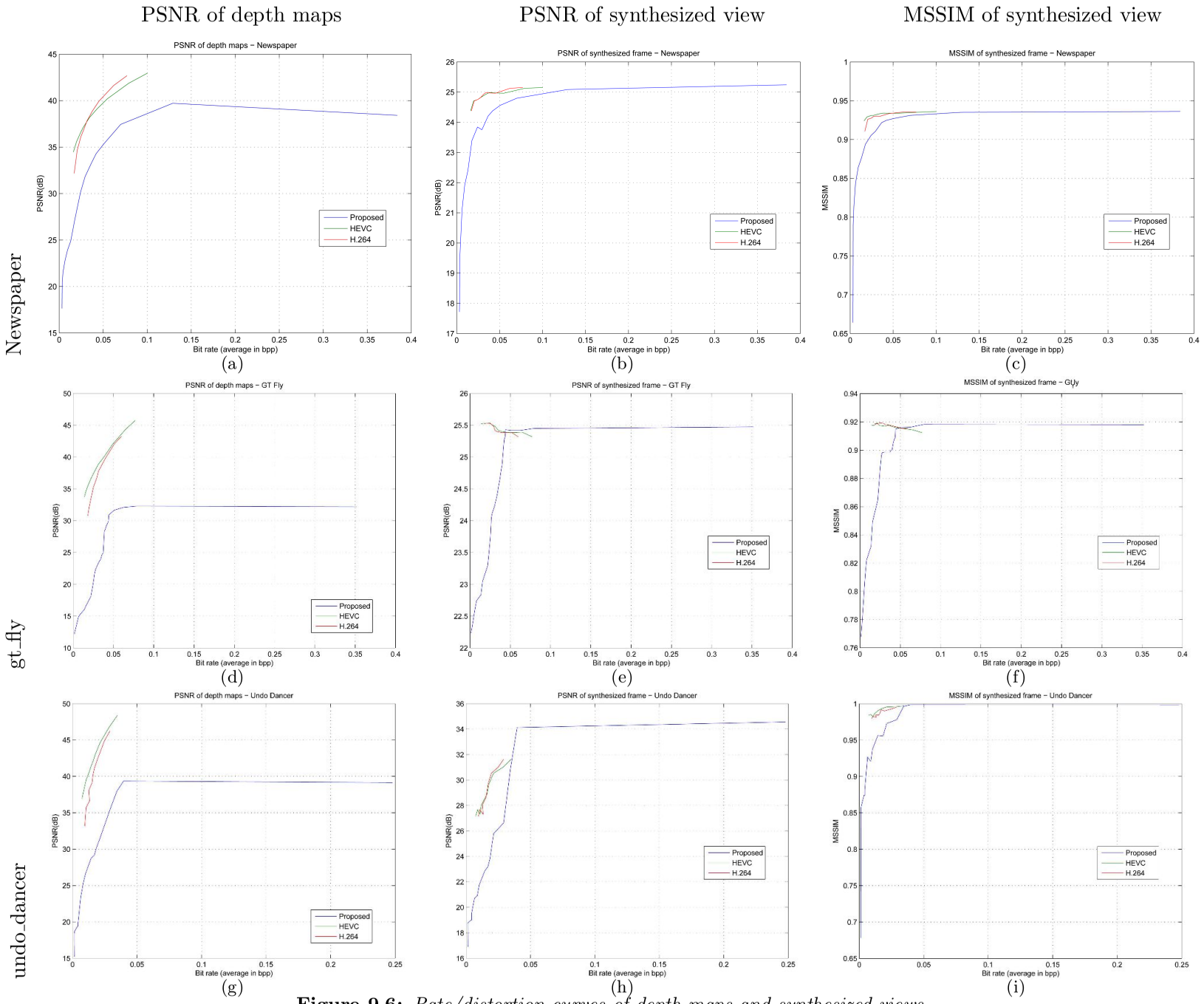
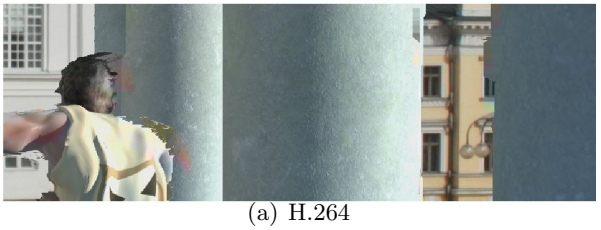
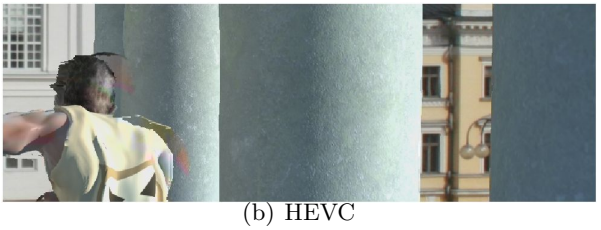


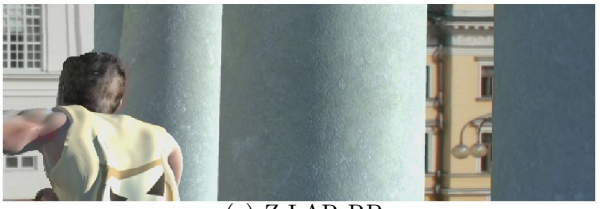
Figure 9.6: *Rate/distortion curves of depth maps and synthesized views.*



(a) H.264



(b) HEVC



(c) Z-LAR-RP

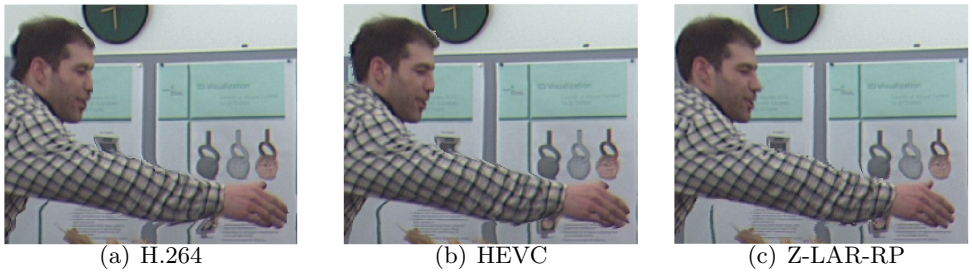
Figure 9.7: *Snapshot of synthesized frame - Undo_Dancer, 0.01bpp.*

(a) H.264

(b) HEVC

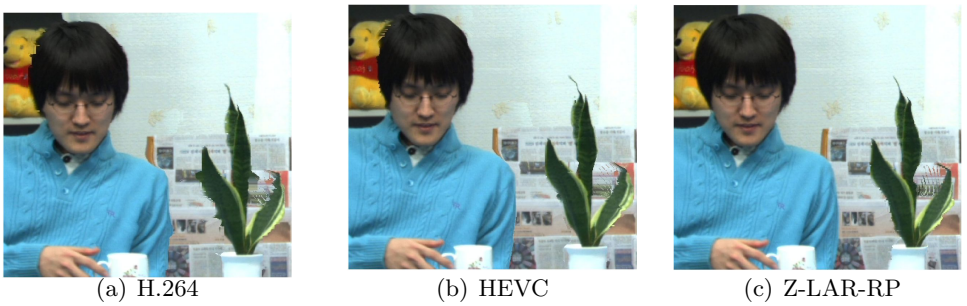
(c) Z-LAR-RP

Figure 9.8: *Snapshot of synthesized frame - GT_Fly, 0.01bpp.*



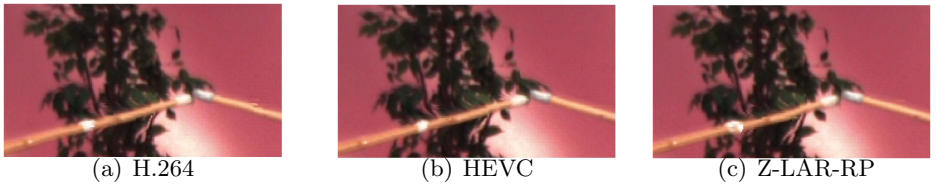
(a) H.264 (b) HEVC (c) Z-LAR-RP

Figure 9.9: *Snapshot of synthesized frame - Book Arrival, 0.02bpp.*



(a) H.264 (b) HEVC (c) Z-LAR-RP

Figure 9.10: *Snapshot of synthesized frame - Newspaper, 0.017bpp.*



(a) H.264 (b) HEVC (c) Z-LAR-RP

Figure 9.11: *Snapshot of synthesized frame - Kendo, 0.01bpp.*

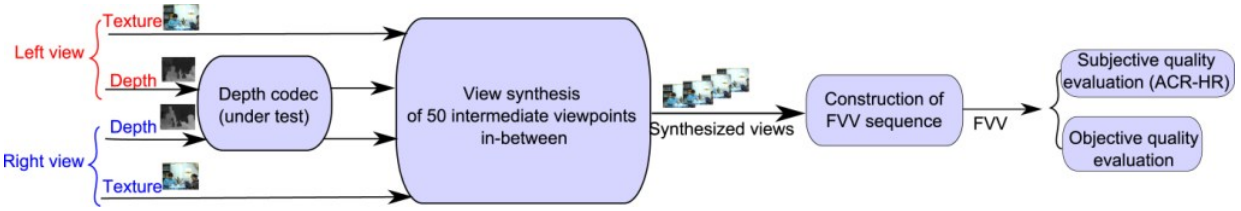
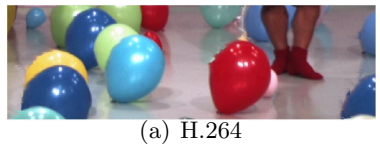
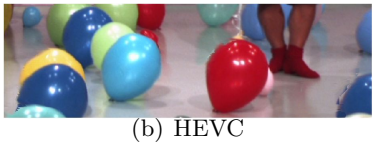


Figure 9.13: *Overview of the experimental protocol.*



(a) H.264



(b) HEVC

(c) Z-LAR-RP

Figure 9.12: *Snapshot of synthesized frame - Balloons, 0.01bpp.*

9.4 Experiment 2: subjective quality assessment

The experimental protocol presented in this section aims at evaluating the impact of depth-compression-related artifacts on the visual quality of the synthesized views. The subjective image quality evaluation test includes the assessment of state-of-the-art codecs. A first subsection presents the experimental protocol used for assessing the compression methods. A second subsection presents and discusses the results.

9.4.1 Experimental protocol

The goal of this experiment is to determine the performances of the Z-LAR-RP coding method, in terms of subjective quality of the resulting synthesized views. So, we consider the impact of depth compression on the quality of views synthesized from the decoded depth maps quality in a FVV context of use. Only depth maps are encoded in order to highlight the impact of depth quantization strategies. Fig. 9.13 depicts the general scheme followed in this experiment. From a given MVD sequence, we consider two different viewpoints and one time t (also referred to as key frames in the following). The associated depth maps are encoded through the depth map codecs under test. From the decoded depth maps, fifty intermediate viewpoints (equally separated) are generated in-between the two considered viewpoints. A sequence of 100 frames (and 10fps) is built from the 50 intermediate virtual frames that simulate a smooth camera motion from left to right and from right to left. This experimental protocol is expected to reveal each coding strategy's

Depth codec	Quantization parameters
H.264 (JM18)	Qp = [{Book Arrival, Balloons, Kendo, Newspaper}{25, 33, 47}, Undo Dancer{25,40,47}, Gt Fly{30,40,47}]
HEVC 6.1	Qp = [{All of the sequences}{34, 45, 50}]
3D-HTM	Qp = [{All of the sequences}{25, 35, 47}]
JPEG2000	0.05bpp, 0.009bpp and 0.005bpp
Z-LAR-RP	Y = {20, 60, 240 }

Table 9.4: *Quantization parameters used in the experiment.*

distortion specificity. Depth coders under test include the Z-LAR-RP, HEVC 6.1 and H.264 (JM 18), 3D-HTM 0.4 (provided by MPEG) and JPEG2000, all in intra coding mode. For H.264, we used the JM 18.4 (Joint Multiview Video Model) software for the Multiview Video Coding (MVC) project of the Joint Video Team (JVT) of the ISO/IEC Moving Pictures Experts Group (MPEG) [jm12]. For JPEG2000, a C++ implementation of the JPEG2000 standard was used [kak12]. In the case of 3D-HTM, inter-view prediction and VSO (View Synthesis Optimization) parameters were enabled. The choice for these methods in this experiment is motivated by the fact that they are reference methods which are usually used as anchors in standardization process. Three test quantization parameters were selected for each depth codec under test according to the visual quality of the rendered views. This procedure was motivated by the need to cover a wide range of categories in the visual quality scale in order to properly define each codec under test. Table 9.4 gives the details of the quantization parameters used in these experiments. Six MVD sequences are used in these experiments: *Book Arrival*, *Newspaper*, *Kendo* and *Balloons* are real scenes; and *GT_Fly* and *Undo_Dancer* are synthetic scenes. Table 9.1 summarizes the features of the sequences. The sequences and the key frames were selected for their availability and amount of depth. Table 9.5 gives the details of the encoded viewpoints and the target viewpoint for the synthesis. The synthesis process is performed through the 3D-HTM 0.4 renderer, that is the view synthesis algorithm used in MPEG 3DV group of standardization at the time of writing this paper. We set the *Blended Mode* parameter of the synthesis algorithm for using the right view only for hole filling instead of carrying out a weigthed average of samples extrapolated from both sides (as done in the MPEG evaluations).

Twenty-seven naive observers participated in the subjective quality evaluation test into two 30-minute sessions. ACR-HR [ITU08] methodology was used to assess 288 FVV sequences, among which were the 96 hereby considered. ACR-HR methodology [ITU08] consists in presenting each stimulus only once to the observers, who are asked to rate the quality of the stimuli relying on a five-level quality scale (5: *Excellent*; 4: *Good*; 3: *Fair*; 2: *Poor*; 1: *Bad*). The reference version of each stimulus is included in the test procedure and rated like any other stimulus. This is reffered to as a "hidden reference condition".The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a Panasonic BT-3DL2550 screen (1920×1080p), and according to ITU-T BT.500 [BT.93]. The stimuli sequences with lower resolution (1024x768) were displayed at the sequence resolution with a grey surrounding to fit the Full HD screen.

Sequence Name	Encoded viewpoints	Frame no.
Book Arrival	$10-6$	33
Newspaper	$2-6$	1
Balloons	$1-5$	1
Kendo	$1-5$	1
GT_Fly	$1-9$	157
Undo_Dancer	$1-9$	250

Table 9.5: *Input and output views of the experiment.*

9.4.2 Results

From the subjective scores obtained with the ACR-HR method, Mean Opinion Scores (MOS) and Differential Mean Opinion Score (DMOS) are computed between each stimulus and its corresponding (hidden) reference. As recommended in VQEG multimedia Test Plan [VQE08], the DMOS are calculated on a per subject per processed stimulus (PS) basis. The corresponding reference version of the stimulus (SRC) was used to calculate an off-set version of the DMOS value for each PS following the expression:

$$DMOS(PS) = MOS(PS) - MOS(SRC) + 5 \tag{9.3}$$

In such conditions, the higher the DMOS, the better the quality of the tested stimulus. The lowest bound is 1 as for MOS values but the highest bound can be higher than 5. If the DMOS value is greater than 5, this means that the stimulus is rated better than its corresponding hidden reference. Such values are considered valid by VQEG [VQE08].
Fig. 9.14 plots the DMOS scores obtained for *Undo Dancer* sequence. In this experimental protocol, the stimuli were not classically selected relying on a list of bit-rates to be evaluated. The stimuli were previously selected by experts based on their subjective visual quality evaluations. For each coding method, the subjective visual quality of the views synthesized from decompressed depth data, at different bit-rates, were first considered by the experts. Then, for each coding method, the experts selected three stimuli corresponding to the categories *Good*, *Fair*, *Poor*. This explains that the obtained curves do not lie in the same bit-rate range. For any coding method, we refer to the highest, the middle and the lowest bit-rates evaluated as $R0$, $R1$ and $R2$ respectively. Fig. 9.14 shows that in two cases (for Z-LAR-RP and for HEVC coding methods), the observers rated the $R2$ better than $R1$ while the visual quality is expected to fall down when the bit-rate decreases. In the case of Z-LAR-RP, for $R2$, the depth maps used to generate the FVV are almost uniform depth maps. This suggests that a uniform depth map, at low bit-rate, induces less annoying artifacts in the FVV sequence. In the case of HEVC, the depth maps for $R2$ contain smooth edges but the structure of the scene is still perceptible. This suggests that some coding strategies induce coding artifacts whose impact on the visual quality of the synthesized views is reduced and preferable, at low bit-rate.

Fig. 9.15 shows the DMOS scores obtained for *Balloons* sequence. For three coding methods, DMOS values are higher than 5 (bold black line in the Figure). Since the reference is rated 5 by definition, this means that the processed sequence is rated with a better quality than its associated hidden reference sequence. This can be explained by the fact that depth estimation errors may be smoothed when processed by some compression methods. This is typically the case around object edges, where depth estimation is prone to errors. Some compression methods for some bit-rates may thus smooth inaccurate estimated depth areas, leading to a better visual quality of synthesis. So, we assume that

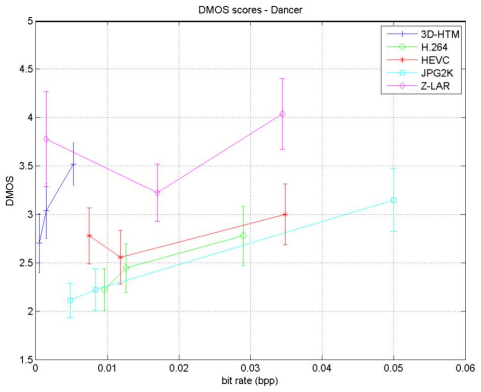


Figure 9.14: *Subjective DMOS over bit-rate - Undo Dancer.*

this phenomenon comes from the impact of coding strategies on inaccurately estimated depth maps. This is a particular phenomenon that can be observed in the context of DIBR-synthesized views.
In Fig. 9.15, the visual quality of $R2$ is also rated better than that of $R1$ with the Z-LAR-RP and the HEVC coding method.

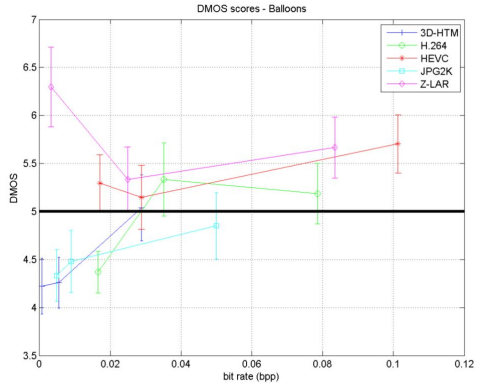


Figure 9.15: *Subjective DMOS over bit-rate - Balloons.*

Figures 9.16, 9.17 plot the DMOS scores for *Book Arrival* and *Newspaper* (the plots for the other sequences are not presented since the results were similar). In these two figures (9.16, 9.17), the Z-LAR-RP coding method also obtains good results in terms of subjective visual quality, at very low bit-rate. These results strengthen the idea that a depth map coding strategy inducing depth fading at low bit rate can enhance the subjective visual quality of the synthesized views. Concerning the performances of the compression methods, they seem to vary according to the video content. This is in accordance with the previous comment regarding the impact of the depth estimation accuracy and of the coding strategy on the visual quality of the synthesized views.

3D-HTM includes VSO which modifies the bit-rate distortion trade-off for encoding side depth maps, considering the impact on a synthesized view. The latter is located on the middle view point between the reference view and the current side view. However, FVV requires to synthesize many in-between views with decoded depth optimized for a

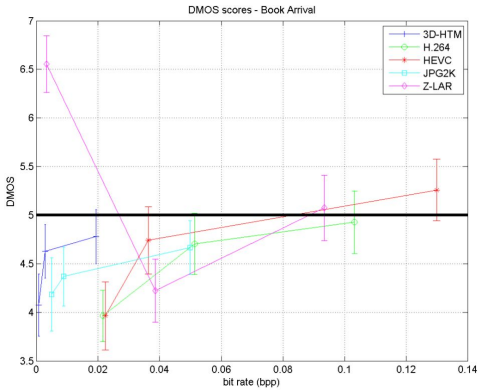


Figure 9.16: *Subjective DMOS over bit-rate - Book Arrival.*

unique view point. This may explain the low performance on Figures 9.16 and 9.17. HEVC outperforms H.264 for all the contents except in the case of *Newspaper*, considering the same range of bit-rate. Similarly, Z-LAR-RP is always rated with the best quality for the considered bit-rate range, except for the cases of *Book Arrival* and *Newspaper*. These examples suggest that a given compression strategy leads to a typical type of distortion that is not perceived or equally accepted depending on the video content. To validate this assumption, an important study on the influence of video contents on compression methods performances is required. We also assume the existence of an impact of MVD sequences features on compression performances.
Finally, an important comment regards the plotted performances of Z-LAR. Except for the cases of *Book Arrival* and *Newspaper*, as previously mentioned, Z-LAR-RP is always rated with the best subjective quality scores. It should be recalled that this compression method relies on a specific strategy which consists in modifying the depth structure of the scene for saving bit-rate. In other words, the lower the bit-rate, the lower the amount of depth in the represented scene. Indeed, in this experiment, the lowest bit-rate corresponds to an almost uniform depth map. And yet, using uniform depth maps for synthesizing new frames amounts to projecting all the reference-colored pixels into the same depth plane. This reduces the errors generally occurring around strong depth discontinuities. Consequently, parallax is significantly reduced in the considered FVV sequences synthesized from these low rate Z-LAR-RP encoded depth maps. For the same reason (uniform depth map), the views rendered from low-bit-rate-Z-LAR-RP encoded depth maps are slightly shifted from the targeted virtual viewpoint, as previously observed in Fig. 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12. As a matter of fact, since Z-LAR-RP tends to shift the scene because of the uniform depth maps, the usual full reference quality metrics penalize the method.Yet, the observers rated the subsequent Z-LAR-RP-sequences with the best scores. The observers may have preferred Z-LAR-RP distortions, that is to say, the lack of parallax, over the compression errors that generally appear around object edges as ringing or “crumbling” artifacts. However, the observers have rated one factor of the 3D QoE: image quality.

9.5 Conclusion

In this chapter, we presented a novel approach for depth coding, relying on LAR method. It takes benefit from a pyramidal profile and allows the encoding of multi-resolution depth maps. The enhancement of low resolution depth maps is performed through the help of a

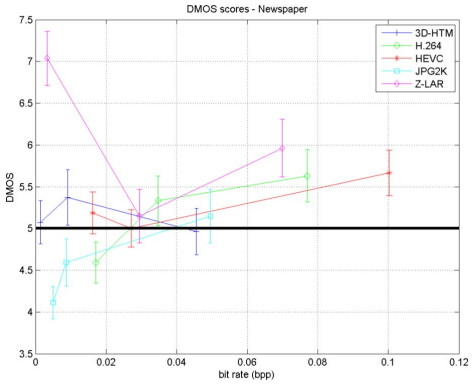


Figure 9.17: *Subjective DMOS over bit-rate - Newspaper.*

region segmentation map obtained from the quad-tree only and improved by the decoded color information. The rate control strategy and the quantization consist in spatially quantizing the depth: the actual depth structure of the scene is modified when the bit-rate decreases, by increasing the homogeneity threshold of the quad-tree partition. The depth map tends to be uniform at very low bit-rates (until 0.003bpp). Altough state-of-the-art coding methods ouperform this novel approach, according to the objective measurements, psycho-visual tests proved that the strategy of Z-LAR-RP enhances the visual quality of the synthesized views, in a FVV context of use. The visual performances achieved thanks to the quantization strategy of Z-LAR-RP show that it may be preferable to transmit less depth depth values than erroneous depth data. The results show that such a depth fading strategy can improve the visual image quality.

Part IV

Relationships between color and depth data

Based on the results of the previous parts and on the literature, we investigated the relationship between texture and depth data. Indeed, when designing a novel coding framework for MVD data, the question of bit rate allocation between the two types of data often raises. This part addresses the analyses we ran in order to study this issue presented in two chapters.
Chapter 10 first investigates bit rate allocation when using H.264/MVC coding method and when relying on PSNR of synthesized frames as a distortion criterion. Then, HEVC coding method is considered using the same experimental protocol. Finally, in Chapter 11, further analyses on sequences' features are proposed in order to explain the variability of the budget allocated to depth according to the sequence.

CHAPTER 10

Bit rate allocation in Multi-view Video Coding

The previous chapters showed that the synthesized views can be distorted by different processes, in particular by the synthesis process and the compression step. Since the quality of the synthesized views is dependent on the accuracy of both texture and depth data, the issue of bit rate allocation between these two types of data must be addressed when designing an MVD coding framework. This chapter questions the bit rate allocation between texture and depth data for encoding MVD data sequences. This question has not been solved yet because not all surveys reckon on a shared framework. The study presented in this chapter includes the compression of MVD data sequences with H.264/MVC and HEVC in intra mode at different bit-rates in order to determine the best bit rate distribution between depth and texture, when based on PSNR measures of the synthesized view. The chapter is organized as follows: Sec. 10.1 presents the reasons that motivated our study on bit-rate allocation in MVD coding context. Sec. 10.2 details the experimental protocol of the study. Sec. 10.3 discusses the results of the study when using H.264/MVC as a MVD coding framework. Sec. 10.4 addresses the results of the study when using HEVC as a MVD coding framework. Finally Sec. 10.5 concludes the chapter.

10.1 Motivations

Since texture and depth information are required for view synthesis in both FTV and 3D-TV, an efficient coding framework should ensure the preservation of essential data. Indeed, previous studies ([MSMW07a],[VYS08]) have shown that coding artifacts on depth data can dramatically influence the quality of the synthesized view. This has been discussed previously in Chapter 5 Chapter 8 and Chapter 9. Depth maps are not natural images. Most of state-of-the-art codecs used for depth maps are based on 2D video codecs that are optimized for human visual perception of color images. Yet, a straightforward idea suggests that being a monochrome signal, depth maps require low bit-rate compared to texture data. Actually, because of its capital role in the view synthesis processing, compression artifact of such data may lead to fatal synthesis errors when generating virtual views. Consequently, a simple but essential question refers to the bit allocation ratio between texture and depth. This rate ratio depends on the targeted application.

Here, we address this question by measuring the Peak Signal to Noise Ratio (PSNR)

scores of intermediate views which need to be generated in contexts of 3D-TV (for rendering on autostereoscopic displays) or of FTV for rendering view points different from those captured by the cameras. In most of the studies, the use of this objective metric is justified by its simplicity and mathematical easiness to deal with such purposes. We investigate the bit-rate allocation issue when based on such a metric.

The appropriate rate ratio that should be used is not clearly stated in the literature: most of the studies do not rely on the same framework. Fehn *et al.* [Feh04] show that being a gray-scale signal, the depth video can be compressed more efficiently than the texture video and recommend using less than 20% of the texture bit-rate for video-plus-depth data format. This recommendation is based on the fact that "*the per-pixel depth information doesn't contain a lot of high frequency components*"[Feh04]. In [LHM+09], the authors proposed an efficient joint texture/depth rate allocation method based on a view synthesis model distortion, for the compression of MVD data. According the the bandwidth constraints, the method delivers the best quantization parameters combination for depth/texture sets that maximizes the rendering quality of a synthesized view in terms of MSE. The proposed model finds optimal ratio between depth bit-rate and texture bit-rate in this paper. However, the optimization depends on the target virtual viewpoint.

Our experiments tried to quantify the appropriate rate ratio between depth and texture data, and then analyze the relationship with the encoded sequence. This study led to the publication of one national [BJM+10] and one international [BJP+11] conference papers in collaboration with INRIA laboratory, in Rennes.

10.2 Protocol

We aim at evaluating the required ratio between depth and texture data relying on the quality of a reconstructed view, in terms of PSNR. In a first case study, H.264/MVC reference software, JMVM 8.0 (Joint Multiview Video Model) is used to encode three views, as a realistic simulation of a 3D-TV use. In a second case study, HEVC 6.1 is used to encode the left and right views. The choice for these methods in this experiment is motivated by the fact that they are reference methods which are usually used as anchor in standardization process. To vary the bit-rate ratio and the total bit-rate, the quantization parameter QP varies from 20 to 44 for both depth and texture coding. The central view predicts the two other views. Then, from the decompressed views, we computed the intermediate view between the central view and the right one, by using the reference software: VSRS, version 3.5, provided by MPEG.

Figure 10.1 illustrates the described protocol. In this figure, "MVC codec" thus refers to either H.264/MVC or HEVC 6.1 depending on the case study. We used two different types of sequences to answer our question: *Ballet* from Microsoft Research, and *Book Arrival* from Fraunhoffer HHI (1024 × 768). This last sequence is a 3DV test material in MPEG. It was acquired through cameras arranged equidistantly along a straight line with a rectified configuration (no gap in-between the cameras and the interaxial distance is 65 mm). On the other hand, *Ballet* was acquired with converging cameras. The considered views are 2, 4 and 6 for *Ballet*, and 6, 8 and 10 for *Book Arrival*. Viewpoint 3 is generated from decoded viewpoints 2 and 4 in the case of *Ballet*. Viewpoint 9 is generated from decoded viewpoints 8 and 10 in the case of *Ballet*. For each couple $(QP_{texture}, QP_{depth})$, the average PSNR score of the synthesized sequence is evaluated, compared to the original acquired view.

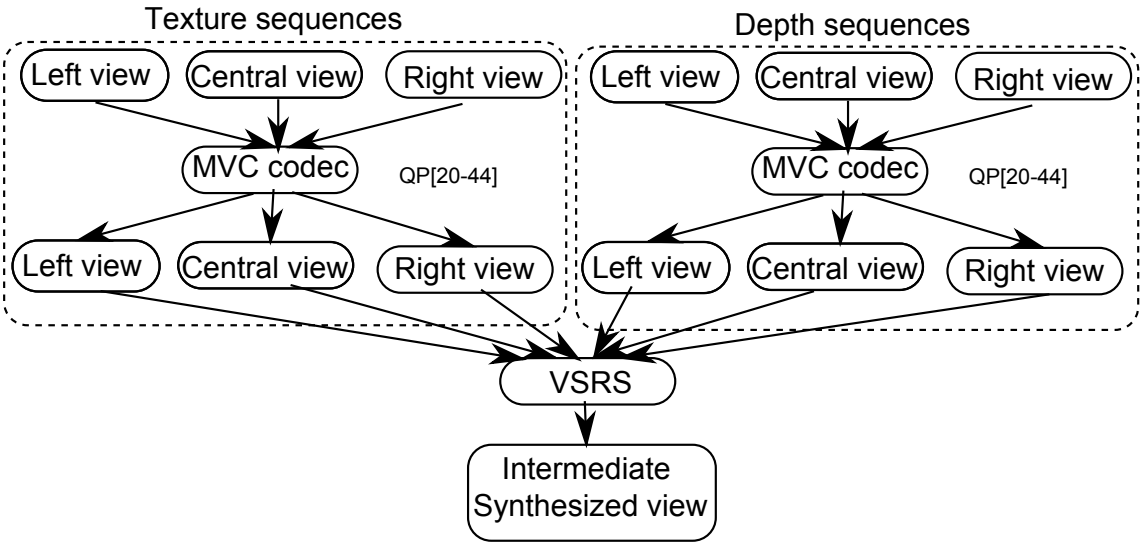
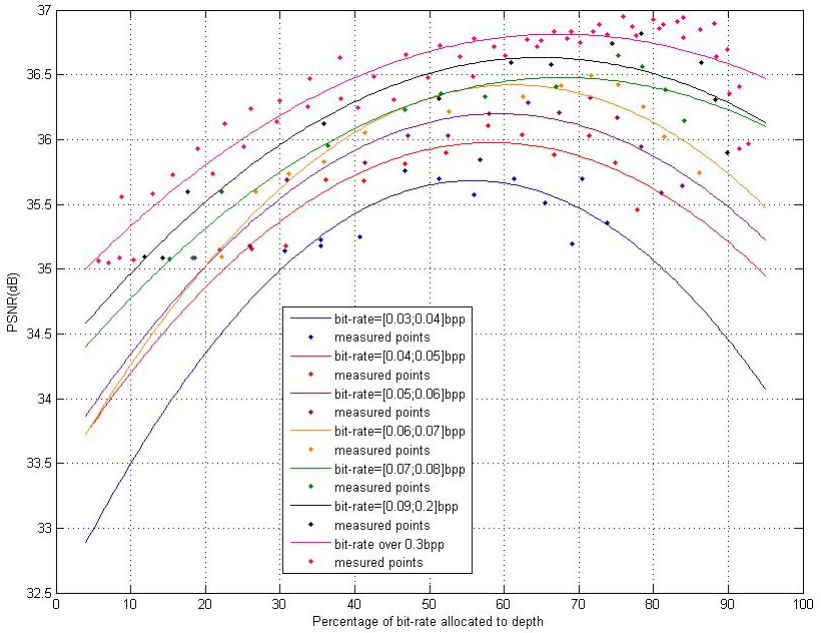


Figure 10.1: *Experimental protocol.*

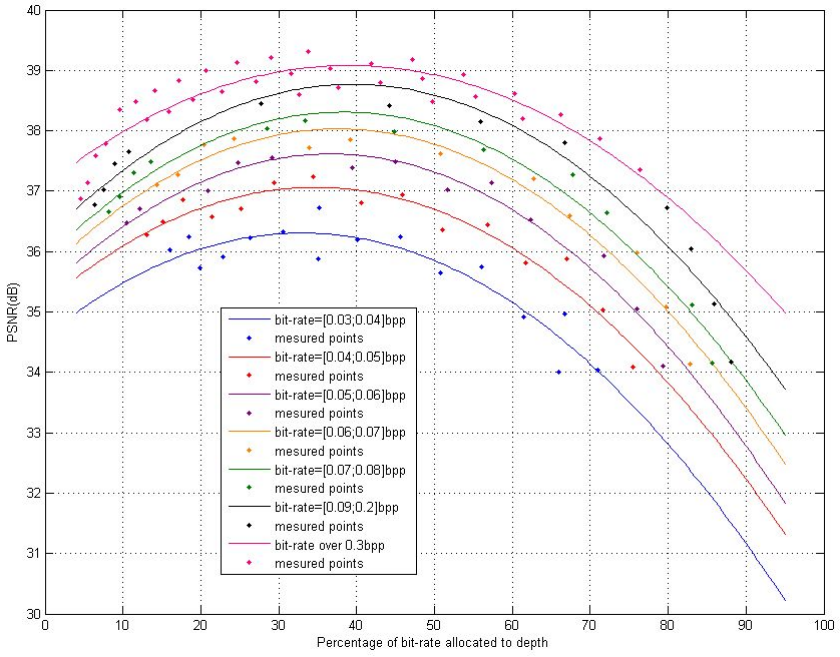
10.3 Bit-rate allocation with H.264/MVC

Figure 10.2 presents the results. Total bit-rate (color+depth) is indicated by colored points. Points with same color belong to same bit-rate range interval. The average PSNR of the synthesized sequence are plotted over the bit-rate percentage assigned to depth. The different curves correspond to interpolation of the measured points for each range of bit-rate. We observe that for a given sequence, no matter the bit-rate, the ratio that provides the best quality is the same: it seems to be around 60% for *Ballet*, and around 40% for *Book Arrival*. This suggests that the required depth information that enables a good reconstruction quality (in terms of PSNR) depends on the content. More precisely, we assume that this percentage depends on the acquisition configuration of the sequence, i.e. the camera baseline. The synthesized sequence from *Book Arrival* may require less elements of depth for the reconstruction because of the linear configuration of the three used cameras. On the contrary, the synthesized sequence from *Ballet*, seems to require an important amount of depth information to ensure a good quality of reconstruction. Correct reconstruction around disoccluded areas may require more reliable depth information depending on the reference camera position. These results are consistent with [LHM+09]: although the authors do not state clearly the appropriate ratio for those two sequences, their rate/distortion curves show that, for example, the bit-rate pair (962kbps for texture, 647kbps for depth), i.e. a percentage of 40% for depth, gives better synthesis quality (in terms of PSNR) for *Book Arrival*. The synthesis conditions are similar to our experiments. On the other hand, in [Feh04], the synthesis conditions involves one single video-plus-depth data: in this case, a very little continuum is supported around the available original view. Since synthesis distortion increases with the distance of the virtual viewpoint, this explains the significant difference with our results.

Figure 10.3 shows particular areas of the synthesized views from the presented experiment. Figures 10.3(a), 10.3(d) and 10.3(g) show that allocating less than 10% of the total bit-rate to depth data induces important damages along the edges. The location of the depth map discontinuities is deeply compromised which leads to errors in the synthesized



(a) PSNR (dB) of synthesized views as a function of rate allocated to depth in percentage of total rate for Ballet



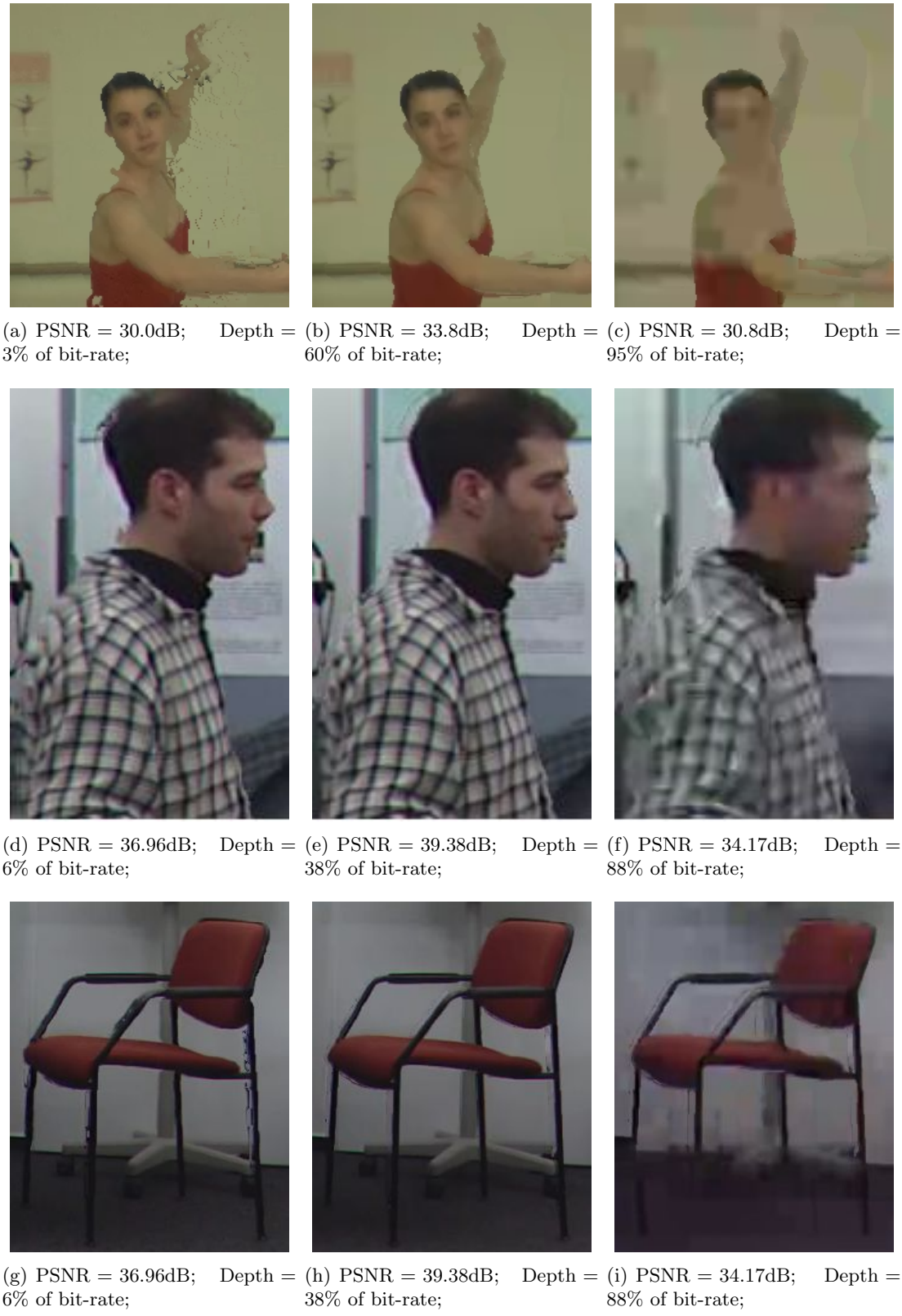
(b) PSNR (dB) of synthesized views as a function of rate allocated to depth in percentage of total rate for Book Arrival

Figure 10.2: *Interpolated rate-distortion curves of synthesized views.*

view. PSNR scores fall down because of the numerous errors along the contours of objects. Figures 10.3(c), 10.3(f) and 10.3(i) suggest that assigning more than 80% of the total bit-rate to depth data preserves the edges of some objects but texture information is lost because of the coarse quantization. Assigning between 40% and 60% of the total

bit-rate to depth data seems to be a good trade-off for the tested sequences, as it can be observed in Figure 10.3(b), 10.3(e) and 10.3(h). PSNR scores and visual quality are both improved compared with the two other presented cases. The depth maps are accurate enough to ensure correct projections and decompressed texture images are good enough to avoid drastic artifacts. The obtained results showed that the best quality of reconstruction by using VSRS may require to assign between 40% and 60% of the total bit-rate to depth data, depending on the available MVD data. The inflection points of the curves obtained in Fig. 10.2 give the percentage of bit-rate allocated to depth data leading to the maximum PSNR. In average, the percentage of bit-rate allocated to depth data leading to the maximum PSNR is 60.5% for *Ballet* and 36.1% for *Book Arrival*.

Those observations are related to H.264/MVC encoding. Using a different encoding framework may lead to a different ratio for depth. The following section proposes a study regarding this issue and will use HEVC coding method instead of H.264/MVC.



(a) PSNR = 30.0dB; Depth = 3% of bit-rate;

(b) PSNR = 33.8dB; Depth = 60% of bit-rate;

(c) PSNR = 30.8dB; Depth = 95% of bit-rate;

(d) PSNR = 36.96dB; Depth = 6% of bit-rate;

(e) PSNR = 39.38dB; Depth = 38% of bit-rate;

(f) PSNR = 34.17dB; Depth = 88% of bit-rate;

(g) PSNR = 36.96dB; Depth = 6% of bit-rate;

(h) PSNR = 39.38dB; Depth = 38% of bit-rate;

(i) PSNR = 34.17dB; Depth = 88% of bit-rate;

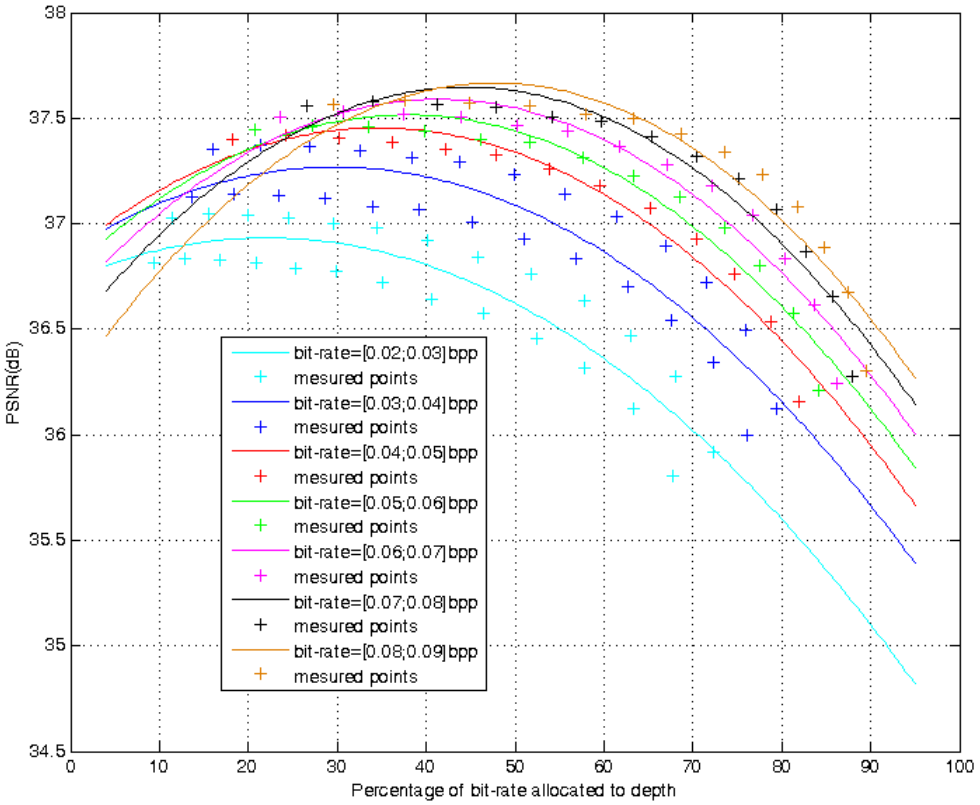
Figure 10.3: *Synthesized images from MVD data, with different bit-rate ratios between texture and depth.*

10.4 Bit-rate allocation with HEVC

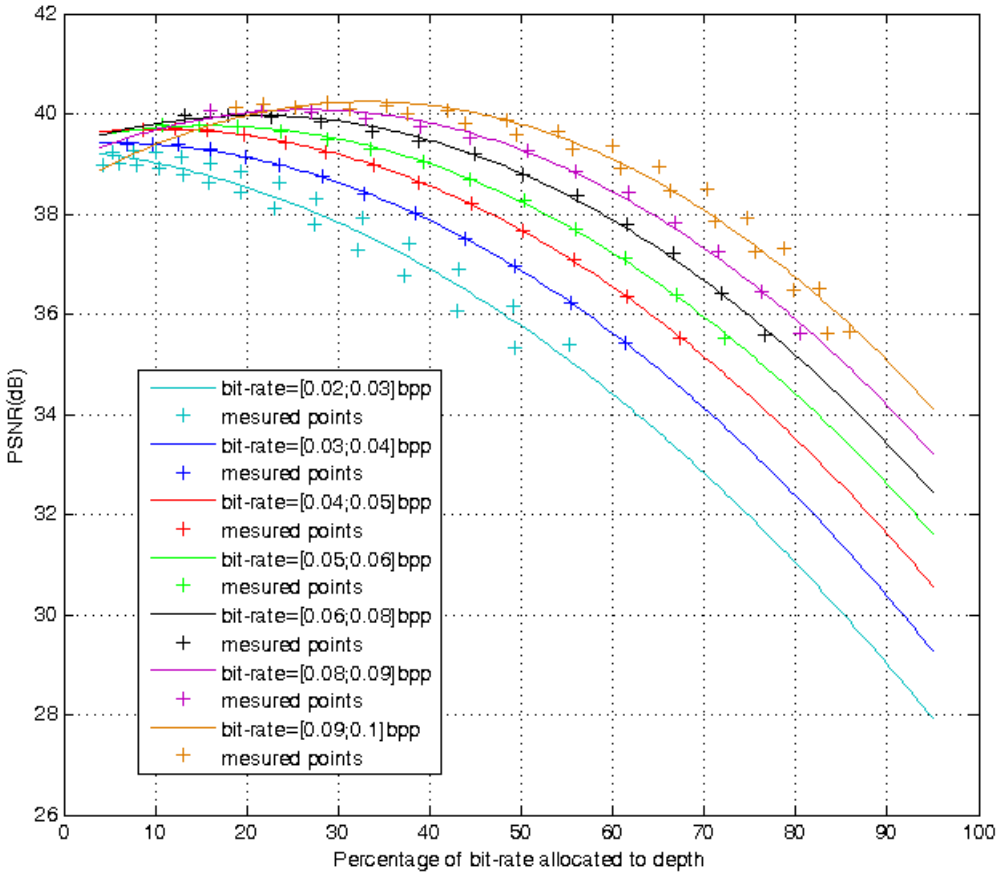
In the previous section, the obtained results showed that the best quality of reconstruction by using VSRS may require to assign between 40% and 60% of the total bit-rate to depth data. Those observations are related to H.264/MVC encoding. In this section, we investigate the bit-rate trade-off using a different coding method: HEVC in intra mode.

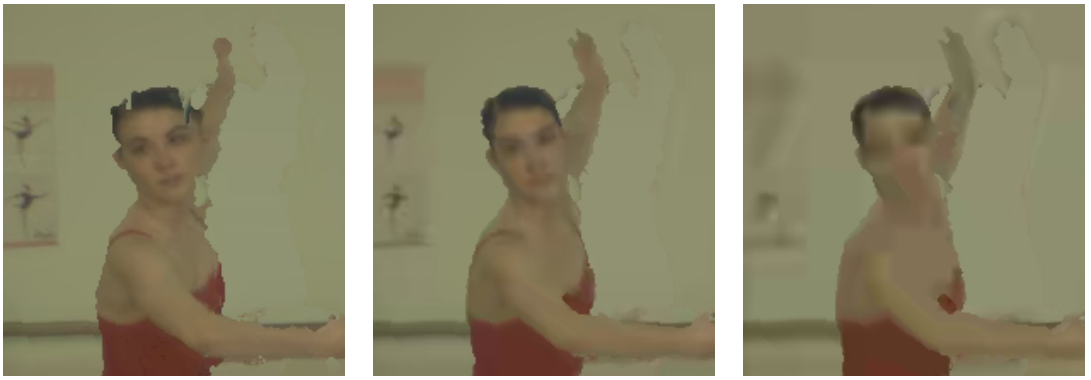
Fig. 10.4 depicts the obtained PSNR curves as a function of the rate allocated to depth in percentage for *Ballet* and *Book Arrival*. In average, the percentage of bit-rate allocated to depth data leading to the maximum PSNR is 27.5% for *Ballet* and 12.2% for *Book Arrival*. Those results are related to HEVC encoding. The obtained ratios are different from those obtained from H.264/MVC encoding, presented in the previous section. These experiments prove that the ratio of bit-rate allocated to depth data in MVD is dependent on the coding strategy.

Fig. 10.5 gives selected snapshots of the resulting synthesized views. This figure can be compared to Fig. 10.3 from the previous experiment. For similar ratios, the visual quality of the synthesized views is different depending on the used encoding method. When 10% of the total bit-rate were not sufficient to render the virtual view of *Book Arrival* properly, when using H.264/MVC. However, it seems to be the best compromise when using HEVC. Fig. 10.5(a), 10.5(b) and 10.5(c) prove that the best compromise, in terms of visual quality lies between 3% and 60%. Fig. 10.5 confirms the objective results in Fig. 10.4.



(a) PSNR (dB) of synthesized views as a function of rate allocated to depth in percentage of total rate for Ballet





(a) PSNR = 36.22dB; Depth = 3% of bit-rate; (b) PSNR = 37.36dB; Depth = 60% of bit-rate; (c) PSNR = 36.3dB; Depth = 89% of bit-rate;

(d) PSNR = 38.99dB; Depth = 6% of bit-rate; (e) PSNR = 38.02dB; Depth = 38% of bit-rate; (f) PSNR = 35.63dB; Depth = 88% of bit-rate;

(g) PSNR = 38.99dB; Depth = 6% of bit-rate; (h) PSNR = 38.02dB; Depth = 38% of bit-rate; (i) PSNR = 35.63dB; Depth = 88% of bit-rate;

Figure 10.5: *Synthesized images from MVD data, with different bit-rate ratios between texture and depth.*

10.5 Conclusion

This chapter presented two studies which aimed at determining the appropriate ratio for joint depth/texture compression, in the MVD framework. The experiments consisted in encoding both texture and depth data by the same compression scheme, varying the ratio between texture and depth information and analyzing the quality of the rendered virtual view.The two experiments aimed at determining the appropriate ratio for joint depth/texture compression, using the H.264/MVC coder in the first case of use, and HEVC in the second case of use. The attributed depth ratio was varied from 2% to nearly 95% and the synthesis of an intermediate view was performed.
A relevant remark regards the observation that for a given sequence (or content) the optimal depth/texture ratio is the same for any total bit-rate; however, the optimal ratio is different depending on the sequence.
The obtained results showed that the best quality of reconstruction by using VSRS require different depth/texture ratios depending both on the content and on the encoding method.

In the next chapter, we propose an analysis of the MVD data and related parameters such as video contents and camera settings to investigate their influence on the best trade-off for the bit-rate allocation between texture and depth data. This next study is meant to help in the conception of tools for automatic bit-rate allocation between texture and depth data.

CHAPTER 11

Impact of features of sequences and bit-rate allocation

The determination of relationships between texture and depth data is useful for the conception of bit-rate allocation strategies, in the context of MVD coding. Based on the previous results, an analysis of different sequence features is proposed to highlight correlations with the best bit-rate allocation. This chapter addresses this question.
Sec. 11.1 reminds the experimental protocol and presents the additional data used for the study of this chapter. Three different aspects are investigated: the entropy of texture and depth data in Sec. 11.2, the discovered areas in the synthesized view in Sec. 11.3 and the color contrast around depth transitions in Sec. 11.4.

11.1 Overview

In this chapter, based on the previous results, we aim at investigating the relationships between texture and depth data. In particular, we assume that the video content, the complexity of depth and the camera settings are related to the bit allocation between depth and texture. The following experiments are in line with this concern.

In total, 11 sequences were included in these tests. For each sequence, we encoded different frames. Then we generated the virtual viewpoints with various baseline distances between the reference viewpoints. Table 11.1 gives the summary of the used material. Table 11.2 summarizes the tested sequences features. The encodings follow the same protocol as described in Chapter 10.2: left and right views (textures and depth maps) are encoded through MVC reference software (JMVM 8.0). Based on the same protocol, the optimal ratio between depth and texture are calculated, with PSNR of the central synthesized viewpoint as an indicator of distortion. Table 11.3 summarizes these results. This table confirms the assumptions raised by the previous experiments since the ratios vary from 16% to 52%: there is a relationship between the content and the required ratio. So the axes to be investigated are the features which differ from one content to another: accuracy of depth map, complexity of depth structure, baseline distance between the reference cameras, features of discovered areas. These aspects are addressed by three analyses in the following: depth maps entropy, baseline distance between the reference cameras, high contrast between background and foreground around the discovered areas.

Sequence Name	Frame no.	Left and Right views	Central view
Ballet	1	0-2	1
	100	0-2	1
Balloons	1	1-3	2
		1-5	3
		3-5	4
	10	1-5	3
	50	1-5	3
	300	1-3	2
		1-5	3
		3-5	4
Book Arrival	1	8-10	9
	99	8-10	9
Breakdancers	1	0-2	1
		0-4	1
			2
			3
		0-6	1
			3
			5
		0-7	1
			4
		1-4	3
		2-6	4
		4-6	5
		4-7	6
	100	0-2	1
Cafe	1	2-4	3
	300	2-4	3
Champagne	1	37-41	39
	300	37-41	39
Kendo	1	1-3	2
		1-5	2
			3
			4
		3-5	4
	300	3-5	4
Pantomime	1	37-41	39
	500	37-41	39
Mobile	1	3-5	4
		3-7	5
		3-7	6
	100	3-5	4
		3-7	5
		3-7	6
	200	3-5	4
		3-7	5
		3-7	6
Lovebird	1	4-8	6
Newspaper	1	2-4	3
		2-6	3
			4
			5
		4-6	5
	2	2-6	4
	10		
	50		
	300		

Table 11.1: *Test material.*

Sequence Name	Characteristics	Depth structure complexity	Camera spacing
Ballet	natural scene, high detail	high	varying, toed-in configuration
Balloons	natural scene, moving cameras, high detail	high	stereo distance, parallel configuration
Book Arrival	natural scene, high detail	high	stereo distance, parallel configuration
Breakdancers	natural scene, high detail	high	varying, toed-in configuration
Cafe	natural scene, medium detail	medium	stereo distance, parallel configuration
Champagne	natural scene, high detail	medium	stereo distance, parallel configuration
Kendo	natural scene, moving cameras, high detail	high	stereo distance, parallel configuration
Lovebirds1	natural scene, natural light, high detail	medium	stereo distance, parallel configuration
Mobile	animation, high detail	simple	stereo distance, parallel configuration
Newspaper	natural scene, high detail	medium	stereo distance, parallel configuration
Pantomime	natural scene, medium detail	high	stereo distance, parallel configuration

Table 11.2: *Features of the tested sequences.*

Sequence Name	Ratio Depth/Texture in %
Ballet	51.6
Balloons	28.21
Book Arrival	31.97
Breakdancers	46.27
Cafe	38.38
Champagne	52.11
Kendo	27.6
Lovebirds	23.58
Mobile	16.57
Newspaper	30.97
Pantomime	19.48

Table 11.3: *Ratio between texture and depth information allowing the minimal distortion in terms of PSNR.*

11.2 Depth maps entropy and texture images entropy

We assume that the ratio that rules the optimal synthesized views in terms of PSNR, is related to the amount of information contained in the original data. In other words, the entropy of depth against the entropy of texture is expected to influence the optimal allocation between depth and texture. Let e_d be the average entropy of the encoded depth maps, for a given content. Let e_t be the average entropy of the encoded texture frames, for the same content. For each tested content, we computed the following ratio:

$$R_e = \frac{e_d}{e_d + e_t} \tag{11.1}$$

Fig. 11.1 plots the computed mean R_e of per sequence against the "optimal" percentage of bit-rate allocated to depth data according to our previous experimental protocol. There is relationship between R_e and the "optimal" percentage of bit-rate allocated to depth data. The correlation coefficient between R_e and the "optimal" percentage of bit-rate allocated to depth data reached 76.95%. These results are understandable because a high entropy value for the depth implies a highly detailed depth structure. If the level of details of depth is higher than that of the texture, the synthesis quality mostly relies on the accuracy of the depth map.
In conclusion, these results suggest that a preliminary analysis of texture and depth entropies can be used as an indicator for automatic bit-rate allocation between these two types of data.

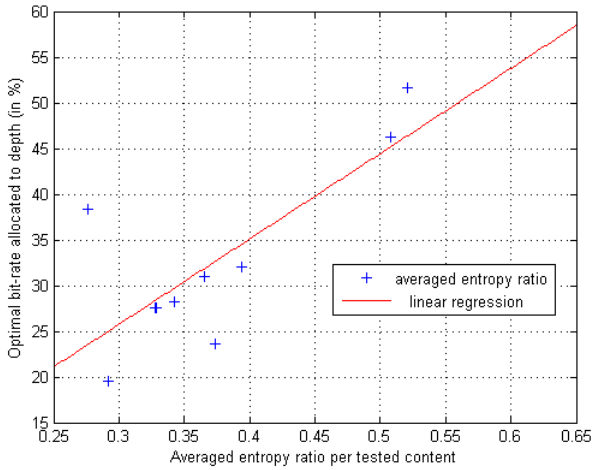


Figure 11.1: *Ratio of entropy between texture and depth data against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR*

11.3 Baseline distance between cameras and discovered areas

We assume that there is a relationship between the structure of the scene depth and the "optimal" percentage of bit-rate allocated to depth data in MVC. According to the depth structure complexity and the baseline distance between the reference cameras, discovered areas in the novel virtual viewpoints are relatively large and difficult to fill-in by the synthesis process. Since the discovered areas are filled in with in-painting methods whose texture estimation quality differs according the used strategy, these areas are prone to perceptible synthesis errors. We aim at evaluating the influence of the discovered areas on the "optimal" percentage of bit-rate allocated to depth data. Let V_r and V_l be the original right and left view, respectively and D_r and D_l be the original right and left depth maps respectively. Let $V_{r\to v}$ the projection of V_r into the target virtual viewpoint, and $V_{l\to v}$ the projection of V_l into the target virtual viewpoint. $V_{r\to v}$ and $V_{l\to v}$ contain undetermined areas that correspond to the discovered areas. $V_{r\to v}$ and $V_{l\to v}$ are used to create logical masks $M_{r\to v}$ and $M_{l\to v}$ defined as:

$$M_{r\to v}(x,y) = \begin{cases} 0 & \text{, if } V_{r\to v}(x,y) \text{ is determined} \\ 1 & \text{, if } V_{r\to v}(x,y) \text{ is not determined} \end{cases} \tag{11.2}$$

$$M_{l\to v}(x,y) = \begin{cases} 0 & \text{, if } V_{l\to v}(x,y) \text{ is determined} \\ 1 & \text{, if } V_{l\to v}(x,y) \text{ is not determined} \end{cases} \tag{11.3}$$

Then we consider the importance I of the discovered areas according to its depth by applying the masks on the respective depth maps as follows:

$$I = \frac{1}{2\times M\times N}\sum_{x=1}^{N}\sum_{y=1}^{M}\left(D_r(x,y)\times M_{r\to v}(x,y) + D_l(x,y)\times M_{l\to v}(x,y)\right) \tag{11.4}$$

where N and M are the width and height of the original image. The score I is computed for each piece of the tested material. In each case, the target virtual point is as indicated in Table 11.1. The results are plotted in Fig. 11.2. This figure shows a linear relation between

the computed importance score I and the "optimal" percentage of bit-rate allocated to depth data. Although the results suggest a relationship between the discovered areas and the "optimal ratio", the virtual viewpoint is not always known at the encoder side. This limits the use of such an indicator for automatic bit-rate control strategies.

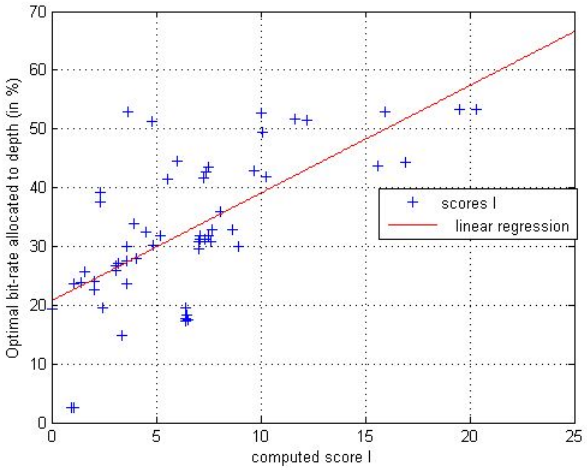


Figure 11.2: *Importance of discovered area against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR*

11.4 High contrast background/foreground areas

We assume that errors occurring after the synthesis process are not only more noticeable when the contrast between background objects and foreground objects is high, but also more penalized by signal-based objective metrics. To investigate this assumption, we consider the strong depth discontinuities (highlighted by an edge detection algorithm) and evaluate the standard deviation of the texture image around these discontinuities. Fig. 11.3 gives an overview of the protocol. This process is applied on right and left views and the final score is the mean of the two obtained measures. Fig. 11.4 gives the plotted scores.

Unexpectedly, the results show that the higher the contrast, the less bit-rate allocated to depth: two main point clouds are distinguishable. The point cloud corresponding to 40-55% of bit-rate allocated to depth belongs to the two toed-in camera configuration sequences. The second cloud corresponds to the parallel camera configuration sequences. So, our assumption is that despite the high contrast around objects contours, the camera configuration (and thus the distance to the virtual view) might reduce the impact of the synthesis distortions.

11.5 Conclusion

This chapter studied the correlation between MVD sequences particular features and optimal bit-rate ratios between texture and depth data as calculated in the previous chapter. This study was meant to provide indicators for designing automatic bit-rate allocation strategies.
The analysis of the MVD data features and related parameters such as video contents and

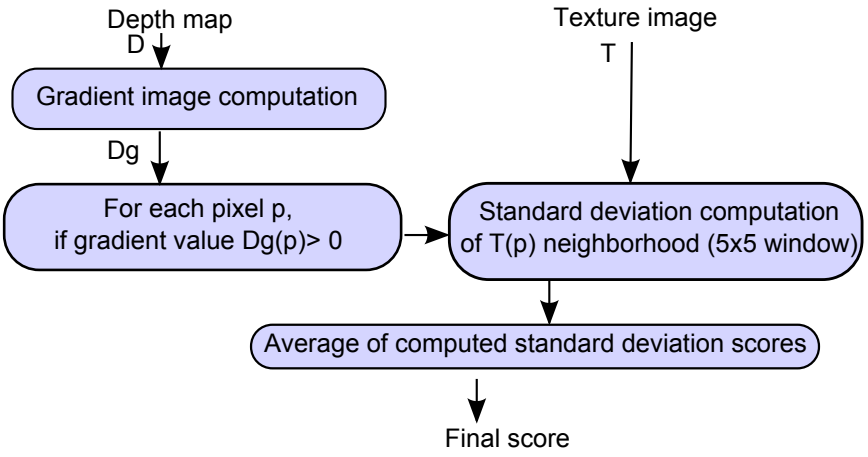


Figure 11.3: *Protocol for the study of the correlation of bit-rate with noticeability of errors in high contrast background/foreground areas.*

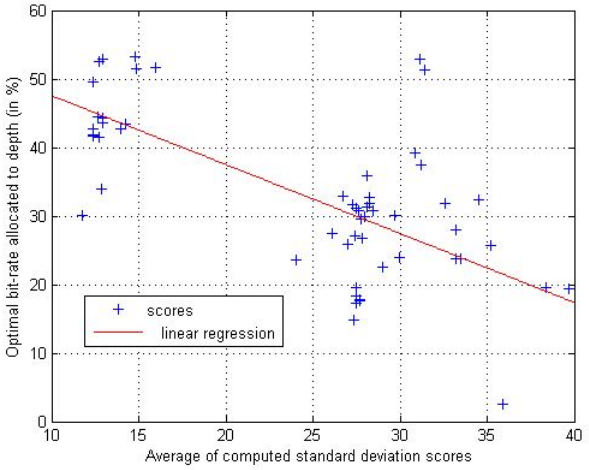


Figure 11.4: *Influence of high contrast background/foreground areas: Average of computed standard deviation scores around gradient pixels of depth maps against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR*

camera settings revealed the existence of their impact on the best trade-off for the bit-rate allocation between texture and depth data. Three different aspects has been investigated in this chapter:

- the relationship between depth maps and texture images entropy and the optimal texture/depth bit-rate ratio,
- the relationship between the discovered areas and the optimal texture/depth bit-rate ratio,
- the relationship between the color contrast around depth transitions and the optimal texture/depth bit-rate ratio.

The results of this study are encouraging because they revealed a relationship with the optimal texture/depth bit-rate ratio and thus the possibility to develop a priori texture/depth

bit-rate allocation methods based on the tested aspects. They also suggest relation between bit-rate allocation and camera configuration of the scene.

Despite its limitations concerning the choice of the distortion indicator (PSNR) that is not perceptually oriented and the media assessment targeting monoscopic viewing, this study highlighted cues for the conception of a priori bit-rate allocation strategies. The new bit-rate allocation strategies might consider a weighted combination of the indicators presented in our study, depending of the used coded.

CHAPTER 12

Conclusion and perspectives

3D Video applications will prosper if we meet the need for high quality content. This requirement is dependent on the ability of 3D systems to provide acceptable synthesized views. DIBR synthesized views are essential for 3D applications such as 3D TV or FTV, so their perceived quality is expected to be at least acceptable for users.
However, there is no adapted tools for evaluating the synthesized views quality. The assessment of such media is not trivial because it involves many human vision factors that are not controlled yet nor understood. The quality evaluation of synthesized views is yet required for the improvement or the performance rating of MVD data compression methods.
This thesis focused on these issues through the analysis of causes of degradations in synthesized views with a view to proposing new tools for 3D Video applications. In the following, we will discuss the contributions and the perspectives regarding each of the three main studied topics: view synthesis related artifacts and the assessment of synthesized views are first addressed; the design of the perception-oriented depth compression scheme is then discussed; finally, we discuss the results and the perspectives following our study on bit-rate allocation between texture and depth data.

Summary of the contributions and perspectives

View synthesis related artifacts and Assessment of synthesized views

Contributions

In this thesis, we investigated the reliability of usual 2D quality assessment methods for the evaluation of synthesized views. Through various DIBR algorithms, novel viewpoints were generated. We highlighted the sources of distortions of the synthesized views. Experimental results showed that the tested subjective methodologies require some adaptations for the assessment of synthesized views, in particular regarding the number of participants.

After studying the assessment methodology of synthesized views, we focused on the related artifacts. First, our analysis showed that they may occur depending on the synthesis strategy. However, we also found out that the artifacts always occurred at specific locations, that is to say around the discovered areas unlike usual distortions such as coding related

artifacts that are scattered over the whole image.

With respect to the measurement of synthesized views with such artifacts, our experiments showed that usual objective metrics fail in rendering higher subjective-correlated quality scores. Our explanation is that most of the usual objective metrics are optimized for coding-related artifacts and thus they do not target the evaluation of the visual rendering quality of objects' edges for example.

We then proposed a new distortion indicator for synthesized frames. The performances of this tool are encouraging but it still needs improvements to be considered as a quality metric.

Perspectives

Following these studies, the design of new objective quality assessment metrics need to be considered. The attempt we proposed for a new objective tool addressing the detection of inconsistent edges is not sufficient. Our observations led us to the assumption that new objective tools evaluating DIBR generated views should involve a registration step, since objects are prone to displacements at the end of the synthesis process. At the moment of writing this thesis, this aspect is investigated in collaboration with IRCCyN (University of Nantes) and University of Roma. The new objective quality assessment tool for synthesized views should be thereafter integrated in coding frameworks for Rate/Distortion optimization strategies and used for evaluating MVD codecs' performances.

It seems of paramount importance to also investigate and standardize a reliable subjective quality assessment protocol for the case of 3D media. However, this will require considerable task forces. So, in the meantime, it can be considered to investigate the reliability of recent 3D-adapted proposed subjective quality assessment protocols through batteries of statistical analyses.

Another challenging effort is that of developing a new subjective quality assessment methodology. Both aspects regarding the presence of new types of artifacts in synthesized views and the new modes brought by simulated stereoscopic viewing conditions should be considered. In particular, from the preliminary experiments conducted during this thesis regarding the stereoscopic viewing conditions, it appeared that relying on the perceived quality of independent views is not sufficient since acceptable artifacts in monoscopic viewing may be annoying in stereoscopic viewing conditions. As well, a new methodology should integrate new requirements, at least for the number of participants. Based on the recent studies recommending the inclusion of new aspects for 3D media assessment, new protocols for the subjective evaluation of 3D media should be studied.

However, the design of new objective metrics and new subjective quality assessment methodologies will require further efforts and important task forces. At the moment of writing this thesis, VQEG is planning activities in order to address these two issues.

Design of a perception-oriented depth compression scheme

Contributions

We studied the impact of depth quantization on the quality of rendered virtual views. A major concern during this thesis was the improvement of the perceived quality of synthesized views. Indeed, based on our previous results, we did not rely on objective metrics for the evaluations of the coding performances, which were only kept as a rough guide.

We proposed two different coding methods, both relying on LAR codec basics. We opted for LAR perception-oriented technique that was basically designed for the compression of still images. Its quad-tree representation allowed a reliable description of the depth structure in the scene. The two extensions we proposed (namely Z-LAR and Z-LAR-RP) only differ at the decoding stage. They both rely on spatial subsampling of the depth map for rate control strategy, by changing the quad-tree representation according to the target bit-rate, instead of quantizing depth values. This is original because at very low bit-rate, the reconstructed depth map is uniform. This method gives priority to the objects' edges quality at the expense of depth feeling: the depth map tends to be uniform while rate decreases. This choice was selected in order to avoid coarse quantization of depth values around the objects' edges, which is known to save bit-rate but increases the artifacts around the rendered objects' edges. At the decoding stage, the first proposed method simply uses neighboring depth information for the prediction of smallest blocks. This prediction is then enhanced through a multi-lateral filter involving the contribution of the corresponding decoded texture image, for preserving the consistency between depth and texture information. The second proposed approach uses a region segmentation method to take benefit from the scalability of the encoding method. In particular, since the region segmentation only requires the decoded quad-tree partition and the corresponding decoded texture view, each level of the pyramidal profile is enhanced based on the knowledge of blocks' belonging to a region. With these two methods, objects may be shifted but no "crumbling" distortions are introduced. As expected the objective metrics rated the proposed schemes as worse than state-of-the-art codecs. This was predictable because our methods change the scene depth structure, because of the evolving quad-tree partition, and because of the averaging of the quad-tree's blocks. However, both the proposed compression schemes showed improvements in terms of visual quality when compared to state-of-the-art codecs such as H.264/AVC and HEVC in intra mode (from our observations).

Perspectives

The exploitation of temporal and inter-view redundancies still needs to be integrated in the proposed LAR extensions. At the moment of writing this thesis, the two proposed depth map compression schemes did not include the exploitation of temporal neither inter view redundancies. They only target still image compression of MVD data. Our first observations also highlighted the importance of depth inter-view coherence when using interpolation-based synthesis algorithms. This suggests to ensure that adjacent decoded depth maps are coherent. These aspects have be treated to enhance the quality of the synthesized video sequences.

As well, another aspect to consider shortly is the evaluation of MVD codec's performances. At the moment of writing this thesis, in the framework of PERSEE project, we are setting

experiments up to evaluate MVD codecs' performances (including Z-LAR-RP's). Based on our preliminary observations, it appears that coding methods that include optimizations based on the quality of the synthesized frames are favored if the coding method includes optimization tools based on the same synthesis algorithm as that used for the final evaluation of codecs. An interesting study would be that of investigating the performances of different coding methods coupled with various synthesis strategies in order to validate our assumptions.

Considering the preliminary observations of MVD codec's performances, the future work for the conception of MVD coding methods should consider the coding/decoding step and the virtual view generation step as closely dependent steps. In particular, the knowledge and the perfect control of the artifacts induced by the encoding method may help for the choice of a synthesis process enabling the best synthesized view perceived quality. This very concern could be presented to MPEG that currently assesses and ranks MVD codecs' performances based on the quality of views synthesized with VSRS, the MPEG view synthesis reference software.

Bit-rate allocation between texture and depth data

Contributions

When designing a novel coding framework for MVD data, the question of bit rate allocation between the two types of data often raises. We thus studied the issue of bit-rate allocation between texture and depth data. Our experiments included the compression of texture views and depth maps at various bit-rates. We combined each texture target bit-rate with each depth map target bit-rate to analyze the relationship between the ratio texture/depth and the objective quality score of synthesized views. Our first study relies on the use of H.264/MVC method. Our second study relies on the use of HEVC. Our results showed that the optimal ratio between depth and texture may differ depending on the encoding method and on the sequence features. An noticeable observation from this study is the following: for a given sequence, the optimal ratio between texture and depth data remains the same for any total target bit-rate, which suggests a relationship between the ratio and the sequence features.

For this reason, a third experiment addressed the role of the sequence features in this ratio, in order to provide cues for a method for a priori setting the ratio without the need for on-line Rate/Distortion optimization. Although the studies investigated several sequences' features that may influence the required ratio between texture and depth information, for allowing the less distortions on the synthesized views, they do not included experiments in the temporal domain. Yet, the amount of movement in the scene, and the depth of the moving objects might have an impact that has not been studied in this work.

Perspectives

New strategies need to be developed including the aspects discussed in this thesis to improve the perceived quality of the synthesized views. Besides, for this very reason, this study should be extended by using more perceptual oriented assessment tools, to increase the robustness of sub-consequent designed a priori bit-rate allocation methods. An interesting aspect that can be explored easily is the appropriateness of the proposed sequence features with respect to subjective quality instead of PSNR (as proposed in this thesis).

Based on this contribution, new coding tools evaluating the sequence features in order to optimize the bit-allocation can be developed shortly.

As for a conclusion, 3D Video involves many challenging issues. This thesis tackled some of them and numerous open questions remain. Important task forces still have to be mobilized in order to improve 3D Video technologies. This thesis meant to highlight the fact that there is a serious need for tools addressing the perceived quality of 3D media and that the success of 3D Video technologies will highly depend on the availability of such tools.

Publications

Book Chapter

1. E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "Visual quality assessment of synthesized views in the context of 3D-TV," in *3DTV System with Depth-Image-Based Rendering: Architectures, Techniques and Challenges*. 2012.

International Journal Paper

1. E. Bosc, R. Pépion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, Nov. 2011.

International Conference Papers

1. E. Bosc, L. Morin, and M. Pressigout, "A content based method for perceptually driven joint color/depth compression," in *Proceedings IS&T/SPIE Electronic Imaging*, pages 8288–82, San Francisco, États-Unis, January 2012.

2. E. Bosc, L. Morin, and M. Pressigout, "An edge-based structural distortion indicator for the quality assessment of 3D synthesized views," in *Proceedings of PCS 2012*, Krakow, Poland, 2012.

3. E. Bosc, M. Koppel, R. Pepion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?," in *ICIP 2011*, Brussels, 2011.

4. E. Bosc, R. Pépion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, L. Morin, and M. Pressigout,"Perceived quality of DIBR-based synthesized views," in *Proceedings of SPIE Optics + Photonics*, San Diego, United States of America, Nov. 2011.

5. E. Bosc, V. Jantet, M. Pressigout, L. Morin, and C. Guillemot, "Bit-rate allocation for multi-view video plus depth," in *Proc. of 3DTV Conference 2011*, Turkey, 2011.

6. E. Bosc, M. Pressigout, and L. Morin, "3D video: new techniques and challenges," in *First Sino-French Workshop on Education and Research collaborations in Information and Communication Technologies, SIFWICT*, Nantes, France, May 2011.

7. E. Bosc, M. Pressigout, and L. Morin, "Focus on visual rendering quality through content-based depth map coding," in *Proceedings of Picture Coding Symposium (PCS)*, Nagoya, Japan, 2010.

Domestique Conference Papers

1. E. Bosc, M. Pressigout, and L. Morin, "Évaluation de la qualité des vues 3D synthétisées," in *Proc. of ORASIS 2011*, Praz-sur-Arly, France, June 2011.

2. E. Bosc, V. Jantet, L. Morin, M. Pressigout, and C. Guillemot, "Vidéo 3D : quel débit pour la profondeur ?," in *Proc. of CORESA 2010*, Lyon, 2010.

PERSEE project internal reports

1. P. Le Callet, V. Ricordel, J. Wang, J. Gautier, C. Guillemot, L. Guillo, O. Le Meur, E. Bosc, L. Morin, M. Cagnazzo, and B. Pesquet-Popescu, "*Livrable D5.2 of the PERSEE project : 2D/3D Codec architecture"*, May 2011.

2. J. Gautier, E. Bosc, and L. Morin, "*Representation and coding of 3D video data"*, Nov. 2010.

3. V. Ricordel, J. Wang, J. Gautier, Le Meur O., and E. Bosc,"*Perceptual modelling for 2D and 3D"*, Nov. 2010.

4. A. Drémeau, C. Guillemot, O. Le Meur, E. Bosc, L. Morin, M. Pressigout, M. Cagnazzo, and E. D'Acunto, "*State of the art in texture analysis and synthesis"*, Oct. 2010.

APPENDIX A

Test MVD sequences

This annex lists the MVD sequences used in the tests of this thesis. This is also one opportunity to thanks all the sequences providers (Microsoft Research, Nagoya University, HHI, GIST, Nokia, ETRI/MPEG Korea Forum).

Ballet

Ballet is a natural scene acquired in a toed-in camera configuration and provided by Microsoft Research. Eight sequences of 100 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Depth maps are computed from stereo vision algorithm. The resolution is 1024×768 pixels and the frame rate is 15 frames per second.

(a) Texture view - Camera 0

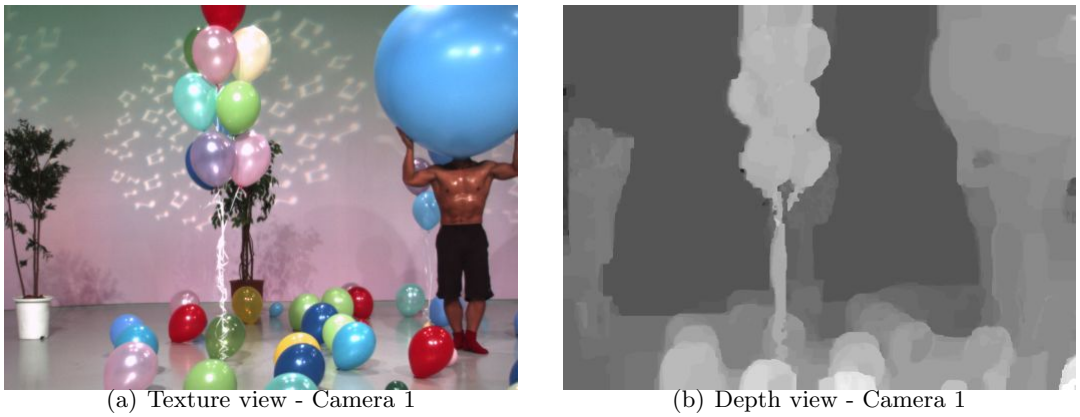
(b) Depth view - Camera 0

Figure A.1: *Ballet sequence.*

Balloons

Balloons is a natural scene acquired in a parallel camera configuration and provided by Nagoya University. Seven sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring

cameras are 5 cm spaced. The resolution is 1024×768 pixels and the frame rate is 30 frames per second.



(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.2: *Balloons sequence.*

Breakdancers

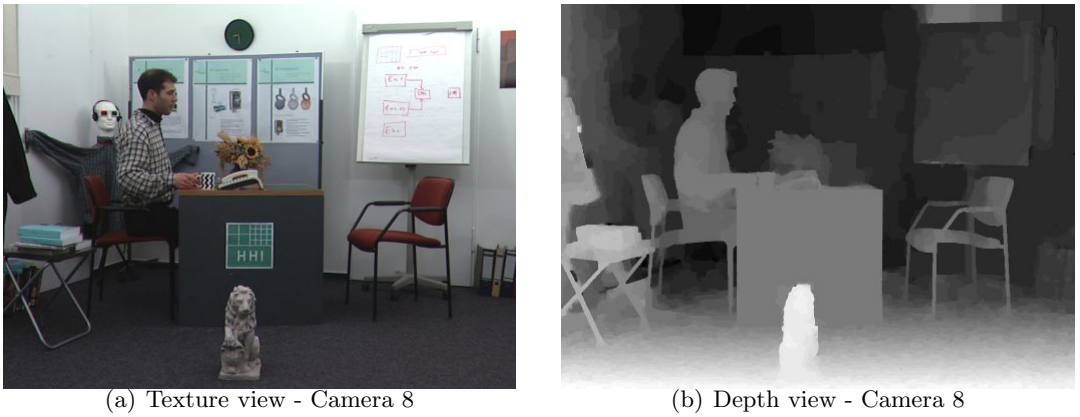
Breakdancers is a natural scene acquired in a toed-in camera configuration and provided by Microsoft Research. Eight sequences of 100 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Depth maps are computed from stereo vision algorithm. The resolution is 1024 × 768 pixels and the frame rate is 15 frames per second.

(a) Texture view - Camera 0 (b) Depth view - Camera 0

Figure A.3: *Breakdancers sequence.*

Book Arrival

Book Arrival is a natural scene acquired in a parallel camera configuration and provided by HHI. Sixteen sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 6.5 cm spaced. The resolution is 1024×768 pixels and the frame rate is 16.67 frames per second.

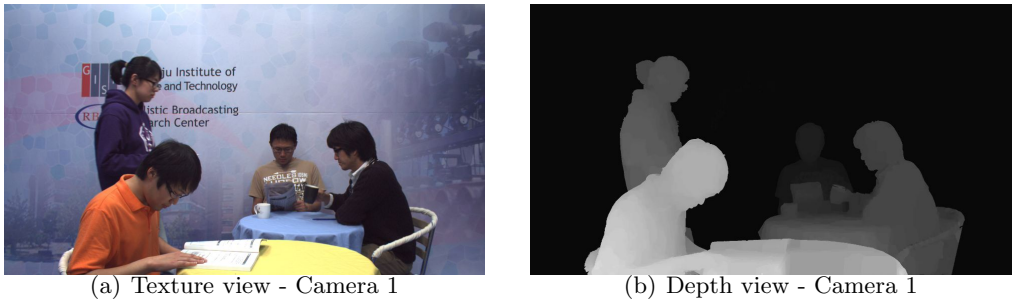


(a) Texture view - Camera 8 (b) Depth view - Camera 8

Figure A.4: *Book Arrival sequence.*

Cafe

Cafe is a natural scene acquired in a parallel camera configuration and provided by GIST. Five sequences of 200 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 6.5 cm spaced. The resolution is 1920×1080 pixels and the frame rate is 30 frames per second.

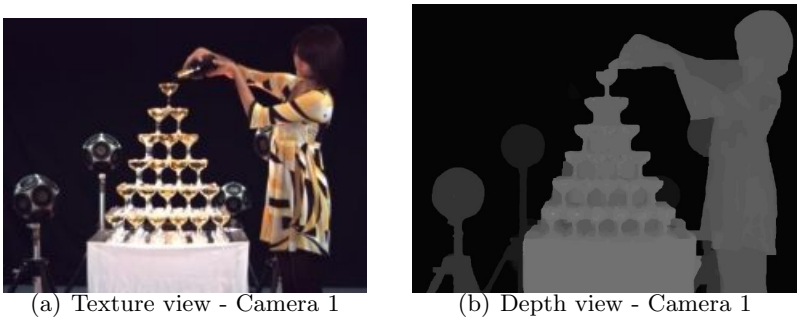


(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.5: *Cafe sequence.*

Champagne

Champagne is a natural scene acquired in a parallel camera configuration and provided by Nagoya University. Eighty sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 5 cm spaced. The resolution is 1280×960 pixels and the frame rate is 30 frames per second.

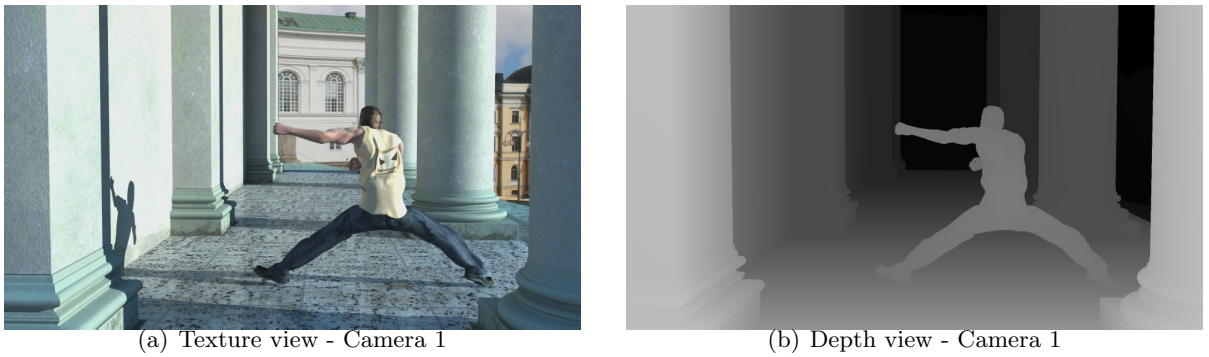


(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.6: *Champagne sequence.*

Undo dancer

Undo dancer is a synthetic scene computer generated in a parallel camera configuration and provided by Nokia. Nine sequences of 250 images are provided together with their associated ground truth depth sequences. Each sequence corresponds to a different viewpoint. The resolution is 1920×1088 pixels and the frame rate is 25 frames per second.

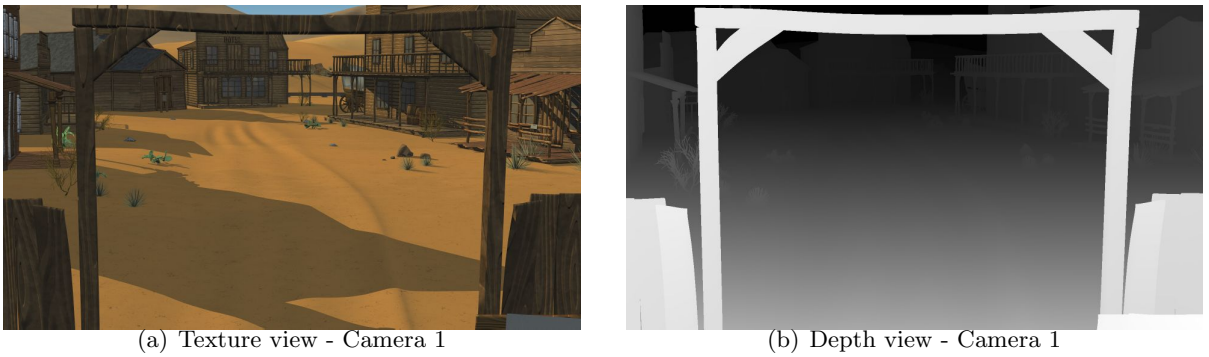


(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.7: *Undo Dancer sequence.*

GT Fly

GT Fly is a synthetic scene computer generated in a parallel camera configuration and provided by Nokia. Nine sequences of 250 images are provided together with their associated ground truth depth sequences. Each sequence corresponds to a different viewpoint. The resolution is 1920×1088 pixels and the frame rate is 25 frames per second.

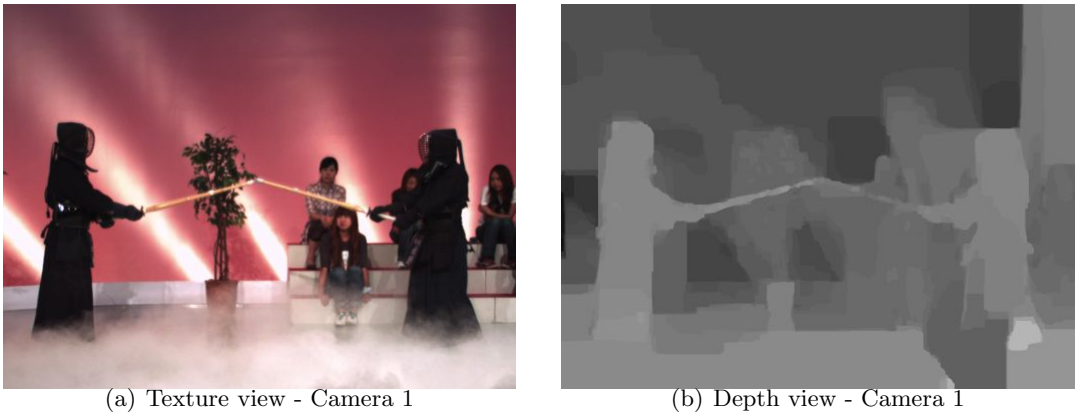


(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.8: *GT Fly sequence.*

Kendo

Kendo is a natural scene acquired in a parallel camera configuration and provided by Nagoya University. Seven sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 5 cm spaced. The resolution is 1024×768 pixels and the frame rate is 30 frames per second.

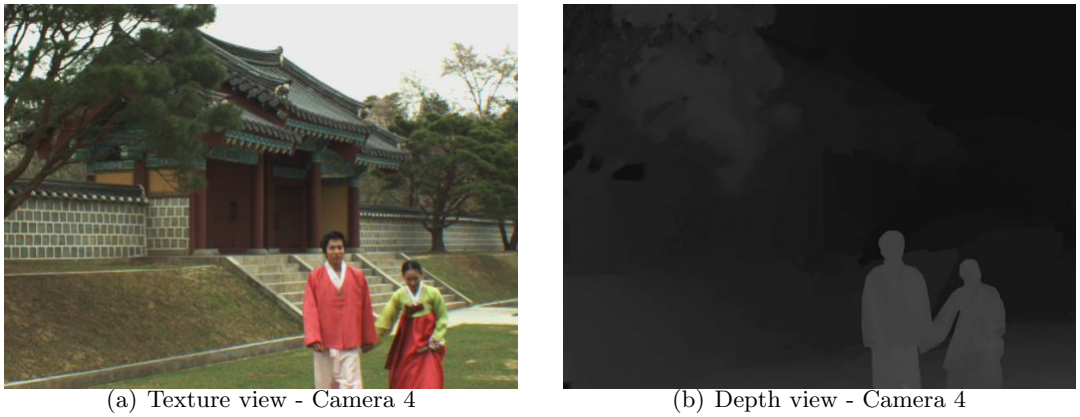


(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.9: *Kendo sequence.*

Lovebirds

Lovebirds is a natural scene acquired in a parallel camera configuration and provided by ETRI/MPEG Korea Forum. Twelve sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 3.5 cm spaced. The resolution is 1024×768 pixels and the frame rate is 30 frames per second.

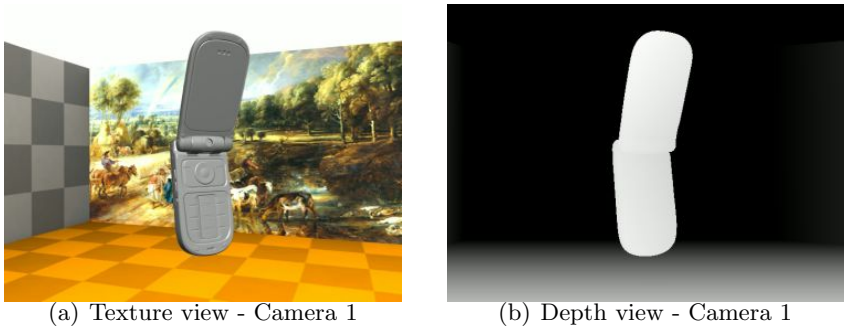


(a) Texture view - Camera 4 (b) Depth view - Camera 4

Figure A.10: *Lovebird sequence.*

Mobile

Mobile is a synthetic scene acquired in a parallel camera configuration and provided by Philips. Three sequences of 200 images are provided together with their associated ground truth depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 5 cm spaced. The resolution is 720×540 pixels and the frame rate is 25 frames per second.



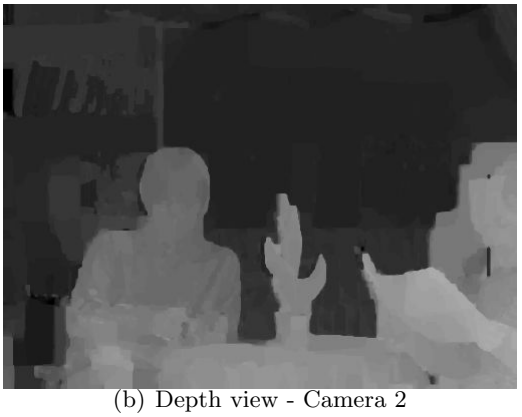
(a) Texture view - Camera 1 (b) Depth view - Camera 1

Figure A.11: *Mobile sequence.*

Newspaper

Newspaper is a natural scene acquired in a parallel camera configuration and provided by GIST. Nine sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 5 cm spaced. The resolution is 1024×768 pixels and the frame rate is 30 frames per second.

(a) Texture view - Camera 2



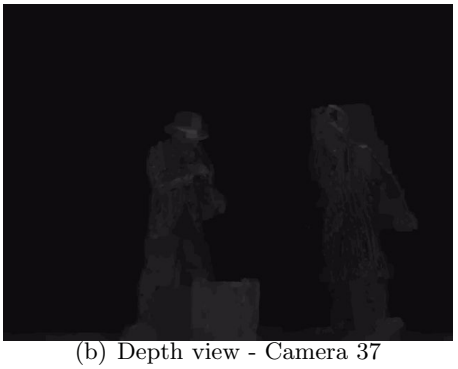
(b) Depth view - Camera 2

Figure A.12: *Newspaper sequence.*

Pantomime

Pantomime is a natural scene acquired in a parallel camera configuration and provided by Nagoya University. Eighty sequences of 300 images are provided together with their associated depth sequences. Each sequence corresponds to a different viewpoint. Acquiring cameras are 5 cm spaced. The resolution is 1280×960 pixels and the frame rate is 30 frames per second.

(a) Texture view - Camera 37



(b) Depth view - Camera 37

Figure A.13: *Pantomime sequence.*

List of Figures

List of Tables

Bibliography

[1112] ISO/IEC JTC 1/SC 29/WG 11. Report of 98th meeting, n12256, December 2012. 28

[AHH+10] P. Aflaki, M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj. Subjective study on compressed asymmetric stereoscopic video. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4021–4024, 2010. 36

[ANR74] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *Computers, IEEE Transactions on*, 100(1):90–93, 1974. 31

[Bab05] M. Babel. Compression d'images avec et sans perte par la méthode LAR locally adaptive resolution. Doctoral thesis, Institut National de Sciences Appliquées de Rennes (INSA), 2005. 86

[BAHG08] M. S Banks, K. Akeley, D. M Hoffman, and A. R Girshick. Consequences of incorrect focus cues in stereo displays. *Journal of the Society for Information Display*, 24(7):7, 2008. 18

[Bar09] M. Barkowsky. *Subjective and Objective Video Quality Measurement in Low-Bitrate Multimedia Scenarios*. Citeseer, 2009. 33, 34

[BCCLC04] S. Battiato, A. Capra, S. Curti, and M. La Cascia. 3D stereoscopic image pairs by depth-map generation. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 124–131, 2004. 16

[BCLC09] M. Barkowsky, R. Cousseau, and P. Le Callet. Influence of depth rendering on the quality of experience for an autostereoscopic display. In *Proceedings of International Workshop on Quality of Multimedia Experience (QoMEX)*, page 6, 2009. 114

[BDR05] M. Babel, O. Déforges, and J. Ronsin. Interleaved s+ p pyramidal decomposition with refined prediction model. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–750, 2005. xix, 88, 90, 108

[BGE+06] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B Akar. Towards compound stereo-video quality metric: a specific encoder-based framework. In *Proc. Southwest Symp. Image Analysis and Interpretation (SSIAI 2006)*, pages 218–222, 2006. 38

[BHG08] A. Boev, D. Hollosi, and A. Gotchev. Classification of stereoscopic artefacts. *Mobile3DTV Project report, available online at http://mobile3dtv.eu/results*, 2008. 32

[BHHB06] M.D. Brotherton, Q. Huynh-Thu, D.S. Hands, and K. Brunnstrom. Subjective multimedia quality assessment. *IEICE Transactions on Fundamentals of Electronics Communications and Computer ScienceES E SERIES A*, 89(11):2920, 2006. x, 54

[BJM+10] E. Bosc, V. Jantet, L. Morin, M. Pressigout, and C. Guillemot. Vidéo 3D : quel débit pour la profondeur ? In *Proc. of CORESA 2010*, Lyon, 2010. 141

[BJP+11] E. Bosc, Vincent Jantet, M. Pressigout, L. Morin, and C. Guillemot. Bit-rate allocation for multi-view video plus depth. In *Proc. of 3DTV Conference 2011*, Turkey, 2011. 141

[BKP+11] E. Bosc, M. Koppel, R. Pepion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet. Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols? In *ICIP 2011*, Brussels, 2011. 53

[Bla89] R. Blake. A neural theory of binocular rivalry. *Psychological review*, 96(1):145, 1989. 13

[BLCCC09] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assessment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008, 2009. 39

[BLCMP12] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout. Visual quality assessment of synthesized views in the context of 3D-TV. In *3DTV System with Depth-Image-Based Rendering: Architectures, Techniques and Challenges*. 2012. 53

[BMP12a] E. Bosc, L. Morin, and M. Pressigout. A content based method for perceptually driven joint color/depth compression. In *Proceedings IS&T/SPIE Electronic Imaging*, pages 8288–82, San Francisco, États-Unis, January 2012. 105

[BMP12b] E. Bosc, L. Morin, and M. Pressigout. An edge-based structural distortion indicator for the quality assessment of 3D synthesized views. In *Proceedings of PCS 2012*, Krakow, Poland, 2012. 53

[BPGA] A. Boev, M. Poikela, A. Gotchev, and A. Aksay. Modelling of the stereoscopic HVS. 59

[BPLC+11a] E. Bosc, R. Pépion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, L. Morin, and M. Pressigout. Perceived quality of DIBR-based synthesized views. In *Proceedings of SPIE Optics + Photonics*, pages 8135–16, San Diego, United States of America, November 2011. 53

[BPLC+11b] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin. Towards a new quality metric for 3-D synthesized view assessment. *IEEE Journal of Selected Topics in Signal Processing*, 5(7):1332–1343, November 2011. 47, 53

[BSB97] L. Bedat, A. Saadane, and D. Barba. Masking effects of perceptual color components on achromatic grating. In *Proc. European Conf. Visual Perception*, 1997. 88

[BT.93] ITU-R BT. *500, Methodology for the subjective assessment of the quality of television pictures*. November, 1993. x, 34, 35, 63, 130

[BT.98] ITU-R BT.2017. Report ITU-R BT.2017 stereoscopic television MPEG-2 Multi-View profile, 1998. 23

[BWS+07] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. von Kopylow. A survey of 3DTV displays: techniques and technologies. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1647–1658, 2007. 16

[CFBLC10] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New requirements of subjective video quality assessment methodologies for 3DTV. In *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010*, Scottsdale, Arizona, U.S.A., 2010. 17, 18, 35, 36

[CFBLC11] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet. New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone. pages 78631O–78631O–13, San Francisco, California, USA, 2011. 15

[CH07] D. M Chandler and S. S Hemami. VSNR: a wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284–2298, 2007. 59

[CJB+12] W. Chen, F. Jérôme, M. Barkowsky, P. Le Callet, et al. Exploration of quality of experience of stereoscopic images: binocular depth. 2012. 36

[CLCB03] M. Carnec, P. Le Callet, and D. Barba. An image quality assessment method based on perception of structural information. In *2003 International Conference on Image Processing*, pages 14–17, Spain, 2003. 39

[CLCM07] P. Campisi, P. Le Callet, and E. Marini. Stereoscopic images assessment. Poznan, Poland, September 2007. 33

[CRM12] P.H. Conze, P. Robert, and L. Morin. Objective view synthesis quality assessment. In SPIE, editor, *Stereoscopic Displays and Applications*, volume 8288 of *Proc. SPIE*, pages 8288–56, San Francisco, USA, January 2012. 39

[CSRK11] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, 57(2):165 –182, June 2011. 37

[CV95] J. E Cutting and P. M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. *Perception of space and motion*, 5:69–117, 1995. 14, 177

[DBBC08] O. Déforges, M. Babel, L. Bédat, and V. Coat. Scalable lossless and lossy image coding based on the RWHaT+ p pyramid and the inter-coefficient classification method. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 185–188, 2008. xix, 88

[DBBR07] O. Deforges, M. Babel, L. Bedat, and J. Ronsin. Color LAR codec: a color image representation and compression scheme based on local resolution adjustment and self-extracting region representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(8):974–987, 2007. 85

[DKG+02] N. Damera-Venkata, T. D Kite, W. S Geisler, B. L Evans, and A. C Bovik. Image quality assessment based on a degradation model. *Image Processing, IEEE Transactions on*, 9(4):636–650, 2002. 59

[DR99a] O. Deforges and J. Ronsin. Locally adaptive resolution method for progressive still image coding. In *Signal Processing and Its Applications, 1999. ISSPA99. Proceedings of the Fifth International Symposium on*, volume 2, pages 825–829 vol.2, 1999. 85

[DR99b] O. Deforges and J. Ronsin. Non-uniform Sub-Sampling using squares elements : a fast still image coding at low Bit-Rate. In *International Picture Coding Symposium PCS 99*, Portland, Oregon, 1999. 85

[DSFK+11] D. De Silva, W. Fernando, H. Kodikaraarachchi, S. Worrall, and A. Kondoz. A depth map Post-Processing framework for 3D-TV systems based on compression artifact analysis. *IEEE Journal of Selected Topics in Signal Processing*, PP(99):1–1, 2011. 105, 109

[DSFW10] D.V.S.X De Silva, W. A.C Fernando, and S. T Worrall. Intra mode selection method for depth maps of 3D video based on rendering distortion modeling. *IEEE Transactions on Consumer Electronics*, 56(4):2735–2740, November 2010. 47

[DTP09] I. Daribo, C. Tillier, and B. Pesquet-Popescu. Motion vector sharing and bit-rate allocation for 3D Video-plus-Depth coding. *EURASIP JASP Special Issue on 3DTV*, page 258920, 2009. 27, 105

[EAP+06] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli. New full-reference quality metrics based on HVS. In *CD-ROM Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, Scottsdale, USA, 2006. 59, 78

[ECW04] T. Ebrahimi, M. Chamik, and S. Winkler. JPEG vs. JPEG2000: an objective comparison of image encoding quality. In *Proc. of SPIE*, volume 5558, pages 300–308, 2004. 38

[EPLC+11] U. Engelke, Y. Pitrey, P. Le Callet, et al. Towards a framework of inter-observer analysis in multimedia quality assessment. 2011. 67, 68

[EWDS+10] E. Ekmekcioglu, S. T. Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz. Depth based perceptual quality assessment for synthesized camera viewpoints. In *Proc. of Second International Conference on User Centric Media, UCMedia 2010*, Palma de Mallorca, September 2010. 39, 40

[Feh04] C. Fehn. Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 93–104, 2004. vii, ix, 15, 19, 38, 47, 60, 61, 79, 141, 142, 178

[FKDB+02] C. Fehn, P. Kauff, M. O De Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton. An evolutionary and optimised approach on 3D-TV. In *Proc. of IBC*, pages 357–365, 2002. 24

[FMG07] M. Flierl, A. Mavlankar, and B. Girod. Motion and disparity compensated coding for multiview video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11):1474–1484, 2007. 24

[G+00] Video Quality Experts Group et al. Final report from the video quality experts group on the validation of objective models of video quality assessment. *VQEG, Mar*, 2000. 67

[Gib50] J. J Gibson. The perception of the visual world. 1950. 12

[Gro] VQEG 3DTV Group. VQEG 3DTV group. `ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/`. xv, 65

[GRP+10] D. B. Graziosi, N. M. M. Rodrigues, C. L. Pagliari, S. M. M. de Faria, E. A. B. da Silva, and M. B. De Carvalho. Compressing depth maps using multiscale recurrent pattern image coding. *Electronics letters*, 46(5):340–341, 2010. 27, 105

[GW] B. Gamber and K. Whiters. History of the stereoscope in 3-D. http://www.bitwise.net/7Eken-bill/stereo.htm. 10

[Han01] J. C Handley. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In *ISand TS PICS Conference*, pages 108–112, 2001. xi, 55

[HB06] M. Haber and H.X. Barnhart. Coefficients of agreement for fixed observers. *Statistical methods in medical research*, 15(3):255, 2006. 67

[HBL+02] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. De Caluwe, S. Kohler, R. H. Koenen, and S. Rihs. PVQM-A perceptual video quality measure. *Signal processing: Image communication*, 17(10):781–798, 2002. 58

[HGS+11] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting*, pages 1–14, March 2011. x, 54

[HPN97] B. G Haskell, A. Puri, and A. N Netravali. *Digital video: an introduction to MPEG-2*. Kluwer Academic Publishers, 1997. 23

[HWD+09] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S. Villette, and A.M. Kondoz. Quality evaluation of color plus depth map-based stereoscopic video. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):304–318, 2009. ix, 53

[HZ03] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Pr, 2003. 45, 46

[IPLW07] J. Ilgner, J. J.H Park, D. Labbé, and M. Westhofen. Using a high-definition stereoscopic video system to teach microscopic surgery. *Stereoscopic Displays and Virtual Reality Systems XIV*, 6490, 2007. 16

[ITU00] ITU. Subjective assessment of stereoscopic television pictures. In *Recommandation ITU-R BT. 1438*. 2000. 35

[ITU08] ITU-T. Subjective video quality assessment methods for multimedia applications. Technical Report Rec. P910, Geneva, 2008. x, 33, 34, 130

[jm12] http://iphome.hhi.de/suehring/tml/, April 2012. 130

[JMFK10] P. Joveluro, H. Malekmohamadi, W. A. Fernando, and A. M. Kondoz. Perceptual video quality metric for 3D video quality assessment. In *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4, 2010. 38

[JQL09] L. Jiangbo, Y. Qiong, and G. Lafruit. Interpolation error as a quality metric for stereo: Robust, or not? In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 977–980, Taipei, Taiwan, April 2009. 40

[JTC07] ISO/IEC JTC1/SC29/WG11. Text of ISO/IEC FDIS 23002-3 representation of auxiliary video and supplemental information. Technical Report Doc. N8768, Marrakech, Morocco, January 2007. 24

[KAF+07] P. Kauff, N. Atzpadin, C. Fehn, M. MÃ¼ller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. *Signal Processing: Image Communication*, 22(2):217–234, February 2007. 15

[kak12] http://www.kakadusoftware.com/, April 2012. 130

[Kau74] L. Kaufman. *Sight and mind: An introduction to visual perception*. Oxford U. Press, 1974. 11

[KCF07] H. Kalva, L. Christodoulou, and B. Furht. Evaluation of 3DTV service using asymmetric view coding based on MPEG-2. 2007. 33

[KNND+10] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand. Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering. In *Proc. IEEE International Conference on Image Processing (ICIP)*, Hong Kong, China, September 2010. x, 47, 60, 61

[LH08] C. Lee and Y. S Ho. View synthesis tools for 3D Video.ISO/IEC JTC1/SC29/WG11 MPEG2008/M15851, October 2008. 46, 177

[LHM+09] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao. Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model. *Signal Processing: Image Communication*, 24(8):666–681, September 2009. 105, 141, 142

[LIH07] M Lambooij, W Ijsselsteijn, and I Heynderickx. Visual discomfort in stereoscopic displays, a review. *SPIE-IS*, 6490, January 2007. 18

[LTL10] P.L. Lai, D. Tian, and P. Lopez. Depth map processing with iterative joint multilteral filtering. *PCS 2010*, 2010. 109

[MdWF06] Y. Morvan, P.H.N. de With, and D. Farin. Platelet-based coding of depth maps for the transmission of multiview images. In *Proceedings of SPIE, Stereoscopic Displays and Applications*, volume 6055, pages 93–100, 2006. 26, 105

[MFdW07] Y. Morvan, D. Farin, and P.H.N. de With. Joint depth/texture bit-allocation for multi-view video compression. In *Proceedings of Picture Coding Symposium (PCS 2007)*, volume 10, page 4349, Lisboa, Portugal, November 2007. 26, 27, 105

[MFY+09] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto. View generation with 3-D warping using depth information for FTV. *Elsevier Signal Processing: Image Communication*, 24:65–72, 2009. ix, 60, 61

[MIS04] M. Meesters, W. Ijsselsteijn, and P. Seuntiens. A survey of perceptual evaluations and requirements of three dimensional TV. *IEEE Transactions on Circuits And Systems for Video Technology*, 14(3):381–391, March 2004. 32

[MLD12] D. Min, J. Lu, and M. N. Do. Depth video enhancement based on weighted mode filtering. *IEEE Transactions on Image Processing*, 21(3):1176–1190, 2012. 109

[MMS+09] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, and T. Wiegand. The effects of multiview depth video compression on multiview rendering. *Singal Processing: Image Communication*, 24(1-2):73–88, 2009. viii, xviii, 105, 106

[MMSW06] P. Merkle, K. Muller, A. Smolic, and T. Wiegand. Efficient compression of multi-view video exploiting inter-view dependencies based on h. 264/MPEG4-AVC. In *Proc. ICME*, pages 9–12, 2006. viii, xviii, 24, 105

[MPE11] MPEG. Call for proposals on 3D video coding technology. Technical report, Geneva, Switzerland, March 2011. 35

[MSD+08] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View synthesis for advanced 3-D video systems. *EURASIP Journal on Image and Video Processing*, 2008. Article ID 438148, 11 pages. ix, 47, 60, 61, 78, 178

[MSD+09] K. Muller, A. Smolic, K. Dix, P. Merkle, and T. Wiegand. Coding and intermediate view synthesis of multiview video plus depth. pages 741–744, November 2009. xviii, 105

[MSMW07a] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand. Multi-view video plus depth representation and coding. In *Proceedings of ICIP*, pages 201–204, 2007. 26, 140

[MSMW07b] P. Merkle, A. Smolic, K. Muller, and T. Wiegand. Efficient compression of multi-view depth data based on MVC. In *3DTV Conference, 2007*, pages 1–4, 2007. viii, 26

[MSMW07c] P. Merkle, A. Smolic, K. Muller, and T. Wiegand. Efficient prediction structures for multiview video coding. *IEEE Transactions on circuits and systems for video technology*, 17(11):1461–1473, 2007. 24

[Mux] Metrix Mux. Metrix mux home page. `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/`. x, 63

[NDK+12] P. Ndjiki-Nya, D. Doshkov, H. Kaprykowsky, F. Zhang, D. Bull, and T. Wiegand. Perception-oriented video coding based on image analysis and completion: A review. *Signal Processing: Image Communication*, 27(6):579–594, July 2012. 2

[NNKD+10] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand. Depth image based rendering with advanced texture synthesis. In *Proc. IEEE International Conference on Multimedia & Expo (ICME)*, Singapore, July 2010. ix, 47, 60, 61

[OEK11] N. Ozbek, G. Ertan, and O. Karakus. Interactive quality assessment for asymmetric coding of 3D video. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pages 1–4. IEEE, May 2011. 36

[OMT+96] T. Okino, H. Murata, K. Taima, T. Iinuma, and K. Oketani. New television with 2D/3D image conversion technologies. *Proceedings of SPIE*, 2653(1):96–103, April 1996. 15

[OS10] R. Olsson and M. Sjostrom. Multiview image coding scheme transformations: artifact characteristics and effects on perceived 3D quality. *STEREOSCOPIC DISPLAYS AND APPLICATIONS XXI*, 2010. 33

[PÓ8] S. Péchard. Qualité d'usage en télévision haute définition: évaluations subjectives et métriques objectives. 2008. xii, 56, 57

[Pal99] S. E Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999. 11

[Pat07] R. Patterson. Human factors of 3-D displays. *Journal of the SID*, 15(11):861–871, 2007. 14, 18, 177

[PBD+08] F. Pasteau, M. Babel, O. Déforges, L. Bédat, et al. Interleaved s+ p scalable coding with inter-coefficient classification methods. 2008. xix, 88

[PBD+10] F. Pasteau, M. Babel, O. Déforges, C. Strauss, L. Bédat, et al. Locally adaptive resolution (LAR) codec. *Recent Advances in Signal Processing*, pages 37–48, 2010. xix, 106

[Per10] C. Perrin. Oeil humain. http://www.biologieenflash.net/animation.php?ref=bio-0029-2, February 2010. 11, 177

[PJ90] N. A Polak and R. Jones. Dynamic interactions between accommodation and convergence. *Biomedical Engineering, IEEE Transactions on*, 37(10):1011–1014, 1990. 12

[PSE+07] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin. On between-coefficient contrast masking of DCT basis functions. In *CD-ROM Proc. of the Third International Workshop on Video Processing and Quality Metrics*, volume 4, 2007. 59

[PW03] M. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In *SPIE Video Communications and Image Processing Conference, Lugano, Switzerland*, 2003. 56

[PW04] M. H Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312–322, 2004. 58

[Res] Video Quality Research. Video quality research. `http://www.its.bldrdoc.gov/vqm/`. 63

[RHFL10] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister. Depth cues in human visual perception and their realization in 3D displays. In *Proc. SPIE*, volume 7690, page 76900B, 2010. 16

[RPLCH10] D. M Rouse, R. Pépion, P. Le Callet, and S. S Hemami. Tradeoffs in subjective testing methods for image and video quality assessment. *Human Vision and Electronic Imaging XV*, 7527:75270F, 2010. 54

[SAB11] M. Solh, G. AlRegib, and J.M. Bauza. 3VQM: a vision-based quality measure for DIBR-based 3D videos. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6, 2011. 40

[SB06] H. R Sheikh and A. C Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430–444, 2006. 59, 78, 113

[SB10] K. Seshadrinathan and A.C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335–350, 2010. 58

[SBdV05] H. R Sheikh, A. C Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *Image Processing, IEEE Transactions on*, 14(12):2117–2128, 2005. 56

[Sch99] C. Schor. The influence of interactions between accommodation and convergence on the lag of accommodation. *Ophthalmic and Physiological Optics*, 19(2):134–150, 1999. 12

[Sch10] B. Schwarz. Mapping the world in 3D. *Nat. Photonics*, 4:429–430, 2010. 15

[SD09] M. Sarkis and K. Diepold. Depth map compression via compressed sensing. In *Proceedings of the 16th IEEE international conference on Image processing*, pages 737–740, 2009. 105

[Seu06] P. Seuntiens. *Visual Experience of 3D TV*. Doctoral thesis, Eindhoven University of Technology, 2006. 35

[SGHS98] J. Shade, S. Gortler, L. He, and R. Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 20

[SMM+06] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand. 3D video and free viewpoint Video–Technologies, applications and MPEG standards. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME06)*, pages 2161–2164, 2006. 19

[SMS+07] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B Akar, G. Triantafyllidis, and A. Koz. Coding algorithms for 3DTV-a survey. *IEEE transactions on circuits and systems for video technology*, 17(11):1606–1620, 2007. vii, 24

[Sou10] G. Sourimant. Depth maps estimation and use for 3DTV. Technical Report RT-0379, INRIA, February 2010. 15

[SS02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 15

[STL04] G. J Sullivan, P. Topiwala, and A. Luthra. The h. 264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions. In *Presented at the SPIE Conference on Applications of Digital Image Processing XXVII Paper No*, volume 5558, page 53, 2004. xviii, 105

[Str11] C. Strauss. Low complexity methods for interpolation and pseudo semantic extraction: applications in the LAR codec. Doctoral thesis, INSA de Rennes, France, 2011. xxi, 118, 119, 120, 179

[SYKH09] Z. M.P Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita. Stereoscopic image quality prediction. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 180–185, 2009. 39

[Tel04] A. Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics, GPU, and Game Tools*, 9(1):23–34, 2004. ix, 60

[TFS+08] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori. Reference softwares for depth estimation and view synthesis. April 2008. xxiii, 46, 78, 79, 92, 111, 178

[TGSM08] A. Tikanmaki, A. Gotchev, A. Smolic, and K. M\Ã¼ller. Quality assessment of 3D video in rate allocation experiments. In *IEEE Int. Symposium on Consumer Electronics (14-16 April, Algarve, Portugal)*, 2008. ix, 53

[VLV96] C. Van, E. Lambrecht, and O. Verscheure. Perceptual quality measure using a Spatio-Temporal model of the human visual system. 1996. 59

[VQE08] VQEG. Final report from the video quality experts group on the validation of objective models of multimedia qualiy assessment, phase 1, 2008. 131

[VYS08] A. Vetro, S. Yea, and A. Smolic. Towards a 3D video format for auto-stereoscopic displays. *Proceedings of the SPIE: Applications of Digital Image Processing XXXI, San Diego, CA, USA*, 2008. 140

[Wan] Y. Wang. Survey of objective video quality measurements. *EMC Corporation Hopkinton, MA*, 1748. 59

[Wan95] B. A. Wandell. *Foundations of vision*. Sinauer Associates, 1995. 11

[Wan06] Y. Wang. Survey of objective video quality measurements. *EMC Corporation Hopkinton, MA*, 1748, 2006. 58

[WB02] Z. Wang and A. C Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, 2002. 56, 78

[WBSS04] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004. 37, 39, 58

[WDK93] A. Woods, T. Docherty, and R. Koch. Image distortions in stereoscopic video systems. 1993. 15, 18

[WF71] H. Wallach and L. Floor. The use of size matching to demonstrate the effectiveness of accommodation and convergence as cues for distance. *Attention, Perception, & Psychophysics*, 10(6):423–428, 1971. 12

[Whe38] C. Wheatstone. Contributions to the physiology of vision. part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical transactions of the Royal Society of London*, 128:371–394, 1838. 9, 11

[Win05] S. Winkler. *Digital video quality: vision models and metrics*. Wiley, 2005. 59

[WLB04] Z. Wang, L. Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, February 2004. 58, 63

[WSBL03] T. Wiegand, G. J Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h. 264/AVC video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):560–576, 2003. 24

[Wu97] X. Wu. Lossless compression of continuous-tone images via context selection, quantization, and modeling. *Image Processing, IEEE Transactions on*, 6(5):656–664, 1997. 90

[WYYJ09] X. Wang, M. Yu, Y. Yang, and G. Jiang. Research on subjective stereoscopic image quality assessment. In *Proc. SPIE*, volume 7255, 2009. 33

[WZ08] Z. F Wang and Z. G Zheng. A region based stereo matching algorithm using cooperative optimization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. 15

[Yam97] H. Yamanoue. The relation between size distortion and shooting conditions for stereoscopic images. *SMPTE journal*, 106(4):225–232, 1997. 15

[Yam06] H. Yamanoue. The differences between toed-in camera configurations and parallel camera configurations in shooting stereoscopic images. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1701–1704, 2006. 15

[YH07] S. U Yoon and Y. S Ho. Multiple color and depth video coding using a hierarchical representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1450–1460, 2007. 25

[YHFK08] S. L. P. Yasakethu, C. Hewage, W. Fernando, and A. Kondoz. Quality analysis for 3D video using 2D video quality models. *Consumer Electronics, IEEE Transactions on*, 54(4):1969–1976, 2008. ix, 53

[YKOH11] K. Yamagishi, L. Karam, J. Okamoto, and T. Hayashi. Subjective characteristics for stereoscopic high definition video. In *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 37 –42, September 2011. 35

[YM94] J. Yang and W. Makous. Spatiotemporal separability in contrast sensitivity. *Vision Research*, 34(19):2569–2576, 1994. 59

[YV09] S. Yea and A. Vetro. View synthesis prediction for multiview video coding. *Signal Processing: Image Communication*, 24(1-2):89–100, 2009. 28

[YW98] M. Yuen and H. R. Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70(3):247–278, 1998. 31, 50

[YWDS+11] S. L. P. Yasakethu, S. T Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz. A compound depth and image quality metric for measuring the effects of packet loss on 3D video. In *Proc. of 17th International Conference on Digital Signal Processing*, Corfu, Greece, July 2011. 40

[YWY+06] Q. Yang, L. Wang, R. Yang, S. Wang, M. Liao, and D. Nister. Real-time global stereo matching using hierarchical belief propagation. In *The British Machine Vision Conference*, pages 989–998, 2006. 15

[YXPW10] J. You, L. Xing, A. Perkis, and X. Wang. Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis. In *Proc. Int. Workshop Video Processing and Quality Metrics, Scottsdale, Arizona, USA*, 2010. ix, 39, 53

[YYED11] X. Yan, Y. Yang, G. Er, and Q. Dai. Depth map generation for 2D-to-3D conversion by limited user inputs and depth propagation. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pages 1–4, May 2011. 15

[ZW09] Z. Zhu and Y. Wang. Perceptual distortion metric for stereo video quality evaluation. *WSEAS TRANSACTIONS*, 5(7), 2009. 35

[ZY10] Y. Zhao and L. Yu. A perceptual metric for evaluating quality of synthesized sequences in 3DV system. In *Proceedings of SPIE*, volume 7744, page 77440X, 2010. 39